2006

Evaluation of a double implanted diffused MOSFET for low power analog applications

Eric J. Basham
San Jose State University

EVALUATION OF A DOUBLE IMPLANTED DIFFUSED MOSFET FOR LOW POWER ANALOG APPLICATIONS

A Thesis

Presented to

The Faculty of the Department of Electrical Engineering

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Eric J. Basham

August 2006

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

APPROVED

For the Department of Electrical Engineering:

We, the undersigned, certify that the thesis of the following student meets the required standards of scholarship, format, and style of the university and the student's graduate degree program for the awarding of the master's degree.

Eric J. Basham
Thesis Author

David Parent (Chair) — Electrical Engineering

Tamara Papalias — Electrical Engineering

Lili He — Electrical Engineering

Wentai Liu — Electrical Engineering
University of California, Santa Cruz

For the University Graduate Committee:

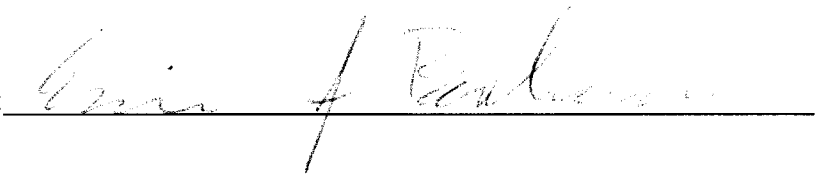
Dean, Division of Graduate Studies

AUTHORIZATION FOR REPRODUCTION

OF MASTER'S THESIS

X I grant permission for the reproduction of this thesis in part or in its entirety without further authorization from me, on the condition that the person or agency requesting reproduction absorbs the cost and provides proper acknowledgment of authorship.

_____ Permission to reproduce this thesis in part or in its entirety must be obtained from me.

Signature of thesis author: 

ABSTRACT

EVALUATION OF A DOUBLE IMPLANTED DIFFUSED MOSFET
FOR LOW POWER ANALOG APPLICATIONS

by Eric J. Basham

The reduced manufacturing and assembly cost, power savings, and increase in reliability are motivation for the move towards integrating analog and digital systems. As a rule, higher levels of integration require additional process complexity and cost. In addition, the bulk of these systems will remain digital and as transistors dimensions fall, device engineering for maximum digital performance compromises analog transistor performance. Asymmetric metal oxide semiconductor (MOS) devices can be included in standard digital complementary metal oxide semiconductor (CMOS) processes with little additional process complexity and display superior analog performance. One class of asymmetric MOS devices – laterally diffused metal oxide semiconductors (LDMOS) – is commonly employed in submicron processes for high voltage applications. A framework which integrates device engineering, model development, testing, and circuit analysis (the g_m/I_d method) is employed to evaluate the low power operation of a LDMOS transistor in a representative CMOS process available at San Jose State University.

ACKNOWLEDGEMENTS

First and foremost I need to thank my long suffering wife. By forming a family, you bring the resources of two individuals together to achieve goals. I had no idea how much my wife would have to contribute when I started on the path to becoming and engineer. She asked only that I enjoy what I was doing, and that has provided me a consistently helpful mantra. A close second set of thanks goes to my advisor David Parent. He had an unusually precise manner of showing me the holes in my argument without discouraging me. It was his posit of a relatively simple question that lead me to investigate this thesis, and along the way, I'd like to think we became friends. Any engineer today works in a highly collaborative environment and so I need to thank my classmates that suffered alongside me – especially Dan Hicks. Dan and I constantly spurred each other to reach further and worked side by side through most of the program. We completed the original proof of concept experiments for this thesis as a humongous class project, and without his help, it probably wouldn't have gone any further.

The companies in Silicon Valley that donated equipment for the microfabrication facility are due hearty thanks as well. Many students benefited from the facilities and I would have had a much different experience if I had not been able to tinker with devices that I fabricated myself. Along that same line, Neil Peters, the fabrication facility support engineer, is due thanks for getting me started and teaching me how to keep things going.

I would like to thank Professor Denis Flandre at Université Catholique de Louvain for posting his excellent thesis on the web. Professor David M. Binkley at the University of North Carolina at Charlotte was kind enough to send a set of lecture notes which greatly assisted the design analysis. Along that line Dr. Wladek Grabinski of Motorola forwarded an excellent extraction tutorial and maintains the EKV model web site.

Table of Contents

List of Tables

List of Figures

Thesis Statement

Asymmetric metal oxide semiconductor (MOS) transistors implemented in a digital process may display superior analog performance compared to standard digital transistors with little additional cost or process complexity. Exploration of these properties through simulation and fabrication is warranted.

Overview of this Thesis

A brief introduction to analog and digital design shows that analog design is a required part of heavily integrated systems, and this trend is likely to accelerate. In the review of analog and digital systems, it becomes clear that scaling for digital performance can have severe impacts on the analog performance characteristics of the transistors and these impacts are poorly predicted by traditional modeling methods. Falling power supplies, part of constant field scaling, necessitate the need for designs which operate with low voltage headroom, but traditional design methodologies fail to facilitate this type of design. A new approach, the "g_m/I_d" or "level of inversion" design methodology, provides several convergent solutions. It is shown that these tools are useful for aiding engineering insight into transistor operation and developing a more intuitive understanding of circuit operation. Due to the closed form nature and the accurate prediction of transistor operation through all levels of inversion, this method also facilitates scaling of analog circuits and computer automated synthesis of analog circuits. Finally, because of the consistent behavior of the models involved, the design approach also allows the comparison of transistor operation across process technologies. It is

suggested that this in turn allows transistor designers direct insight into the predicted analog circuit performance space that the circuits built with these transistors will encompass.

Integration of transistors optimized for analog design is an expensive option, but necessary for integration of high voltage applications, and high voltage applications are becoming more common with thinner gate oxides resulting from scaling efforts and integrating more diverse systems in a single chip. Operation and design of power transistors are reviewed. Due to the unique structure of power transistors they are shown to be an alternative to standard transistors for very low current, high gain circuits. To test this hypothesis, an experimental transistor is simulated, fabricated, and then tested following the g_m/I_d design methodology presented in the thesis. Results show that the experimental transistor shows behavior worth further investigation as a result of its heritage.

Introduction

The foundations of digital computation were worked out before the transistor was invented. Seminal work by Turing, Von Neumann, and others showed that simple switches could be used to design a universal computation machine able to solve a wide array of problems. This is one reason why the use of transistors as binary switches is so widespread. The mathematical formulations are well developed, and by treating the transistors as switches the complications of the physical behavior of the transistors can be simplified and thus abstracted, and the design cycle lends itself to optimization by the devices it produces (i.e., computers can make better computers).

Analog electronics on the other hand takes into account the continuous operation of the transistor both at the input and output, along a continuous time interval. Thus, many additional parameters are important to consider during circuit design. Many of these parameters are intimately entwined with the physical operation of the transistor. As a case in point, there are one or two cutting edge complementary metal oxide semiconductor (CMOS) manufacturing processes for digital design which are mostly a function of the achievable photolithographic resolution, while there are several of types of substrate materials and dozens of types of fabrication processes for analog design as a function of power requirements, noise tolerance, absolute gain, frequency and the inclusion of precision passive components in line with the fabrication process. These factors combine to make abstraction of analog design challenging.

If this is the case why invest effort in integrating analog functionality at all? The current attitude in industry is if you can do it in digital – don't do analog [1]. However,

implementing this design concept is not as strait forward as one might assume because interfacing with the "outside" world is inherently a continuous operation. For example, in a signal processing circuit the incoming signals are continuous and output signals are continuous as well. The circuits performing analog to digital conversion and digital to analog conversion are themselves analog circuits. So while analog circuit design requires significantly more effort per function and unit area, and specialty expertise in analog circuit design, having transistors that perform well in the analog regime will always be part of any highly integrated circuit. The move toward higher levels of integration embodied in the system on a chip revolution (SoC) only exacerbates this demand [2]. Lastly, it has been shown that analog signal processing has a lower bound of energy consumption than digital signal processing [3], so significant gains may be achieved by adopting analog methods. As battery technology has not followed an equivalent of a Moore's law, batteries remain a significant portion of any portable system's composition and mass. This means that as the functionality (operations per unit time) of system increases, the power per function must decrease to maintain a constant power source size. This is a primary driver for low power design efforts.

The Impact of Scaling on Analog Design

One of the most successful ways of reducing power consumption and increasing operational frequency has been the application of scaling methods (Table 1). Scaling methods were originally developed for digital processes and fall into three main

categories: optical shrinking, constant field scaling and constant voltage scaling. An optical shrink typically requires the adjustment of gate oxide and backgate doping and additional process parameters, but does not require a redesign of the circuits or a new layout. Constant field scaling seeks to maintain the field across the transistor as constant. Constant voltage scaling scales parameters to maintain the supply voltage constant. All of these scaling methods improve the performance of digital transistors through improvements in the power delay product as seen in Table 1 [4].

Table 1. Scaling Methods for Digital Design. The Variable Held Constant is the Column Heading, α is the Scaling Factor. Thus a Constant Field Scaling of α=2 Would Reduce the Gate Length by ½, and the Reduce the Power 75%.

Parameter	Symbol	Constant Field Scaling	Constant Voltage Scaling	Constant Voltage Scaling with Velocity Saturation
Gate length	L	$1/\alpha$	$1/\alpha$	$1/\alpha$
Gate Width	W	$1/\alpha$	$1/\alpha$	$1/\alpha$
Field	E	1	α	α
Oxide Thickness	t_{ox}	$1/\alpha$	$1/\alpha$	$1/\alpha$
Substrate Doping	N_α	α^2	α^2	α^2
Gate Capacitance	C_g	$1/\alpha$	$1/\alpha$	$1/\alpha$
Oxide Capacitance	C_{ox}	α	α	α
Transit Time	T_r	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha$
Transit Frequency	Fr	α	α^2	α
Voltage	V	$1/\alpha$	1	1
Current	I	$1/\alpha^2$	α	1
Power	P	$1/\alpha^2$	α	1
Pdp	PDt	$1/\alpha^3$	$1/\alpha$	$1/\alpha$

This effect has resulted in the famous "Moore's Law" - unbroken since it's inception in 1965. Unfortunately these methods typically have negative impacts on analog circuit scaling. For example an 80% optical shrink would result in capacitances shrinking 64% which may result in circuit instability (since the capacitances set poles and zeros in the feedback loops) [5]. Constant voltage scaling, heavily employed until the 0.5μm node, was relatively strait forward to implement, but significant effort in process design and manufacturing was required to deal with the increased electric fields which resulted from this type of scaling [6]. During this period there was a significant movement from hybrid bipolar CMOS processes (BiCMOS) for mixed signal integration to CMOS-only analog capable processes due to the significant cost savings involved. However, because of the process design and fabrication effort involved, electric field scaling gained popularity to further scaling progress. This was in part due to the power dissipation figure of merit for digital processes, as defined by [7]:

$$P = \frac{1}{2} f C_{Load} (V_{Supply})^2 + V_{Supply} I_{Leak}$$

Where P is the total power consumed, f is the frequency of operation of the circuit, C_{Load} is the switching load, V_{Supply} is the bias voltage supply and I_{Leak} is the average leakage current during operation. Two important points are immediately obvious. First, significant power savings are achievable by lowering the supply voltage. Second, while the average leakage current reaches a maximum during the switching process, leakage during the off state can significantly impact the power consumption of the circuit. This is especially true in the case where switches do not operate near full duty cycle.

Thus designing transistors for falling power supplies, increasing switching conductance and lowered standby leakage currents is a clear goal for digital transistor design.

Digital transistor operation design goals have very important impacts upon the operation of analog circuits. In analog circuits, reducing the power supply (which reduces the available swing) actually increases power consumption if the performance is held equal. To understand the impact of scaling power supply, according to [8], the SINAD (signal to noise and distortion ratio) is introduced to show that even though similar specifications are maintained, the power consumption actually increases at technology nodes below 0.25 – 0.35μm. This trend is predicted to continue for ultra deep submicron (UDSM) processes currently under development [9]. An alternate figure of merit introduced for analog to digital converters (ADCs) according to [10] as:

$$FoM = \frac{DynamicRange[2^{2n}] \cdot Bandwidth}{PowerDissapation}$$

using contemporary scaling predictions to confirm this concern below the 0.25μm node. Reference [10] also contains a table oriented for analog design performance for the interested reader. Further degradation of analog transistor behavior is only exacerbated by the lack of V_t scaling to reduce standby leakage current as fixed threshold voltages reduce the total available signal swing as the power supply falls. In this case it is clear that as the power supply falls, even though the g_m of the transistor improves (mostly due to the increase of C_{ox}), an overall increase in power consumption is required to maintain the same frequency and SINAD performance. This is well in excess of the classical signal to noise ratio (SNR) limit of $P=8kT*SNR*f_{signal}$. This situation could be mediated

by using higher power supply voltages, in which case the gate oxide would have to be thicker to tolerate higher gate voltages as well, again incorporating another set of design compromises.

Scaling and the power supply reductions which result thereof are only part of the process redesign effort. Generally, scaling rules offer a guide which process engineers modify to achieve optimum digital transistor operation. While nothing prevents analog designers from using transistors optimized for digital operation when integrating analog and digital circuits on the same chip, the device engineering solutions driven by scaling demands are not complementary. This is easily illustrated with fundamental circuit design elements. The CMOS inverter in digital design and the common source amplifier (capacitively or resistively loaded) in analog design (Figure 1) are simple circuit elements but display behavioral properties which can be used to illustrate the contradiction in design approaches. In digital design, the critical parameters are saturation current and frequency of operation, described by the propagation delay equation and the equation for saturation current. The analog design space is a little more complex.

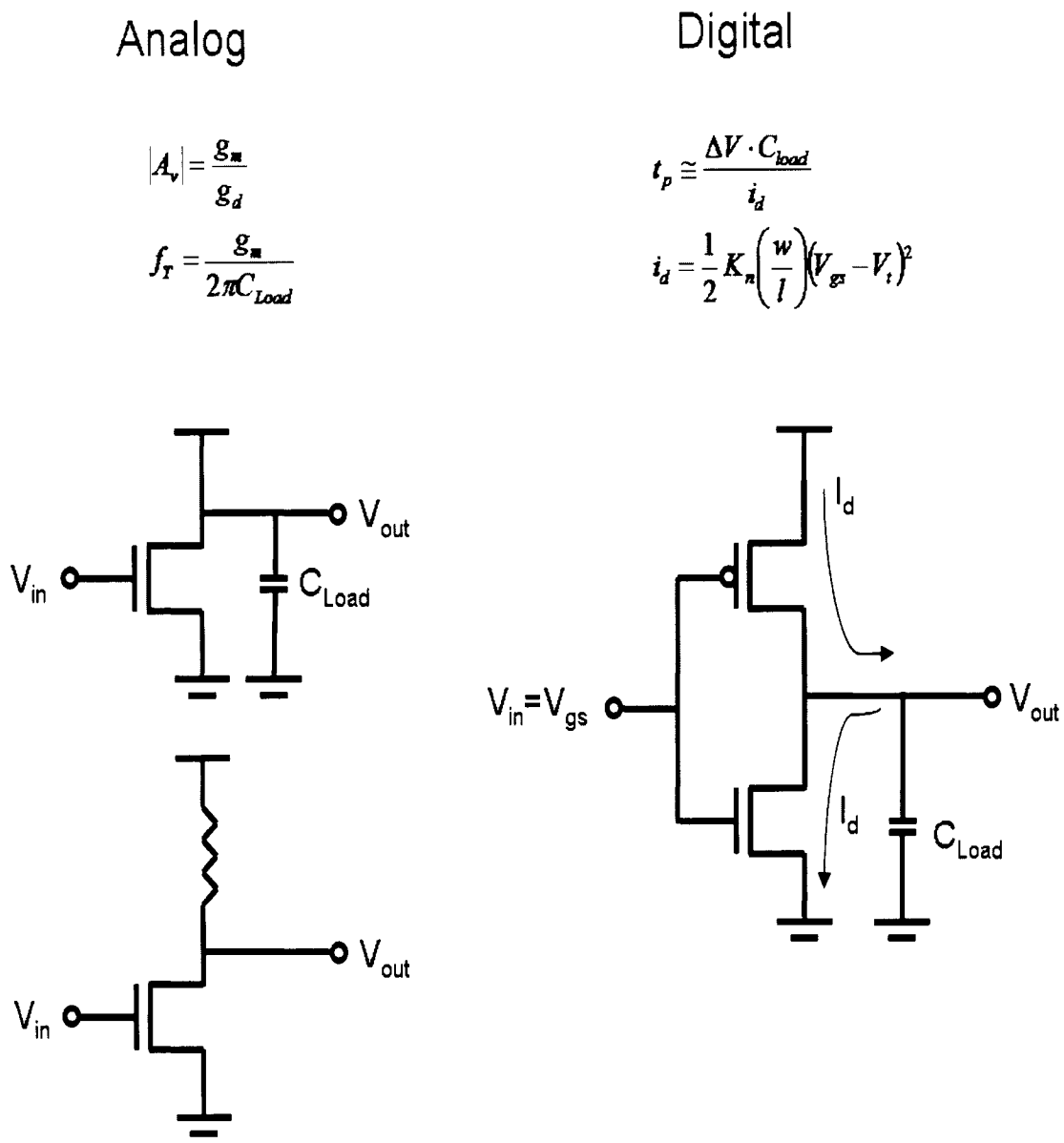


Figure 1. Fundamental Circuit Elements

The classical method of analysis involves evaluating interrelated equations for gain (A_v) and transition frequency (f_T) and then comparing these to an analog design space represented as a tradeoff between power consumption, frequency and SNR [3]. Generalized analysis can become unreasonably complex for large circuits or rely heavily

on design intuition and "rules of thumb" gained by experience. This is one reason that novice designers often rely on a "design by sweep" approach to analog circuit design, and it was observed that [1]:

Simulation time * experience = constant

In point of fact, Razavi suggests analog design space has eight parameters where any two are traded off against each other [11]. One can see how Pelgrom would observe that analog designers generally fail to provide a concise and complete set of technology requirements [12]. Again, this is partly due to the fact that each analog circuit may have different process technology requirements as a function of the application of the circuit.

Because of these fundamentally different design boundaries, the effects of optimizing process technology for digital design on analog performance are dramatic. Analog transistors are required to have a high output resistance and an intrinsic transistor gain greater than 1 (and typically much larger than that) to allow gain bandwidth tradeoff in amplifier design. Since amplifier gain is the product of intrinsic gain and output resistance, lowering either of these parameters will reduce the usefulness of the analog circuitry. Addition of circuitry to compensate for the lack of available intrinsic gain does not offer a direct solution since the addition of gain stages reduces the available input range or output range and increases noise and power consumption. Furthermore, the added design complexity only demands more design effort in an already demanding design environment. In contrast, a low output resistance can actually improve the performance of a digital circuit because higher drain currents charge and discharge the source and drain parasitic capacitances faster. This increased drive current can source or

sink larger gate currents in downstream logic blocks, as shown in Figure 1 (source/drain current arrows). For analog operation, since:

$$R_{ds} = \frac{V_{ea} L}{I_{ds}}$$

increasing I_{ds} and reducing L even at a fixed V_{ea} results in circuit gain reduction.

Controlling leakage current and increasing switching conductance further degrade the analog operation of transistors [13]. While there are gains in intrinsic transistor gain (individual transistor) and transition frequency for analog operation, these are often offset by loss of amplifier gain (amplifier circuits made of transistors), increase in power consumption or loss of SNR. In addition, constant field scaling simultaneously increases device transconductance and reduces all capacitances, thus shifting the circuit operating points. Scaling has long been observed to have detrimental impacts on gain even though transconductance increases for analog circuits [14]. In contrast, constant field scaling a factor of α results in a factor α^3 increase in the power-delay product for digital circuits, as shown in Figure 1. According to Boser and Murmann, analog-digital codesign or "digitally assisted analog" circuit design approaches are warranted since digital metrics have outstripped analog performance gains [15]. However, a solid understanding of analog design effort at the core of these circuits is still necessary.

This situation is complicated as processes scale to dimensions lower than the 0.35μm technology node. It would be useful to have a complete picture of how scaling impacts analog design, but understanding how scaling optimized for digital operations affects analog design is problematic. At each process technology node, the physical effects with

the greatest impact on analog design shift subtly. This is because the scaling rules are used as a starting point and significant engineering effort goes into modification of the scaling rules to maximize digital transistor operation. Typically, each of the direct applications of scaling has some negative effect on analog operation of transistors. This is in bulk due to the tradeoffs necessary to continuously reduce the transistor dimensions.

For example, a major concern in digital deep submicron circuits is the off-state leakage current. This may be mediated by adjusting the threshold voltages upwards with channel doping implants. Thus, threshold voltages do not scale with power supply reduction in digital circuits because scaling the V_t linearly would increase off-state leakage through drain induced barrier lowering (DIBL). This increase in V_t with respect to the supply voltage further reduces the available voltage swing of analog circuits. Another example is alteration of V_{gs}-V_t to increase switching speed. Scaling theory shows that scaling a transistor reduces the switching speed and parasitic capacitances associated with the source and drain area. Further improvements are possible by increasing V_{gs}-V_t. Since transition frequency is proportional to V_{gs}-V_t (because the transistor moves quickly into the saturation region) and g_m is proportional to 1/ V_{gs}-V_t, engineering for higher speed is an anti goal to high gain transistors according to application of the square law presented according to [16]:

$$g_m = \frac{2I_D}{V_{gs} - V_t}$$

Increasing the drive current further reduces achievable gain as a function of power, and will be discussed later. While g_m does increase, it increases less as a function of the bias

current as the transistor moves deeper into strong inversion operation, impacting low power performance. Since the relationships for both the digital case and the analog case have interrelated variables, a clearer method of understanding transistor design effects will be discussed below.

Doping implants in the channel region to adjust V_t, and retrograde implants to reduce gate induced drain leakage (GIDL), also impact analog circuits by increasing mismatch error. The total mismatch can be estimated by the squares of the individual processing variation according to the general form:

$$\sigma^2_{total} = \sigma^2_{Well} + \sigma^2_{VtAdjust}$$

While mismatch is much less relevant in digital circuit design it is crucial in analog design. Greater detail on matching is provided according to [12].

Additionally, as length scales down even further, smaller transistors become more likely to reach velocity saturation. This means the g_m does not increase with smaller L and the device moves from subthreshold conduction directly to saturation operation according to [12]:

$$g_m = \frac{w}{l} C_{ox} \mu (V_{gs} - V_t) \Rightarrow w C_{ox} v_{sat}$$

Subthreshold devices are diffusion mediated, where strong inversion (SI) devices are drift mediated. In SI, an inversion layer provides a means of carrier flow between the source and drain. Below the V_t of the device, no inversion layer is assumed to be present, thus carrier flow is blocked by a potential barrier equal to the bulk potential.

Since subthreshold MOS devices are diffusion devices, it is not possible to achieve velocity saturation. This is a critical analog design fact. Square law design methods would fail to predict operation with any accuracy in this design regime. Even if the transistor is maintained in below velocity saturation, the output drain conductance is likely to fall as a function of channel length modulation effects as a percentage of the total channel width. Even though V_{ea} (early voltage per unit length) increases slightly because of the higher substrate doping, at the 0.35μm node the output conductance becomes a function of the drain current because of the electrostatic pull from the drain voltage on the inversion layer. The impact of output conductance on analog designs can be severe since amplifier gain $A_v = -(g_m/g_d) = -(g_m/I_d)V_{ea}L$. Indeed, by applying gradual channel approximations and using hydrodynamic transistor modeling to analog scaling according to [17] it is demonstrated that the non ideal effects will impact analog design at the 0.1μm – 0.2μm node resulting in a falling gain bandwidth product (GBW) as a function of scaling. The author notes that source drain engineering for analog applications will play a critical role in maintaining analog performance at this point. This remains an active research area and a point of concern [13].

Fortunately, the drive towards smaller transistor dimensions has some significant advantages for analog designers due to the process technology (as opposed to the transistor technology) required. The denser transistor packing drives the need for more metal lines and thus more compact analog circuits. This is especially advantageous for the integration of high metal layer metal insulator metal (MIM) capacitors and inductors far from the substrate with isolating ground shields. Typical epi-substrate with a low-

resistivity bulk silicon, which is used to improve yield and suppress latch-up, causes significant high frequency losses [18] so moving metal layers far from the substrate can be helpful to achieve higher frequencies. There is also a move towards using lightly doped substrates instead of epitaxially grown layers to reduce substrate noise and reduce the self inductive effects of metal line wires [19]. In addition, good layout practice can reduce substrate noise coupling performance in low-ohmic, epi-type substrates, while layout techniques are much less effective for high-ohmic substrates [20]. Therefore, careful layout, typical in analog designs, becomes an asset when using low-ohmic, epi-type substrates. Since interconnect delay actually increases with reduced scale [21], reducing the capacitive and resistive parasitics and the number of layers is of great concern to digital designers, and thus likely to receive significant research focus. Multiple poly lines for dual V_t transistors employed for power saving and leakage current reduction are also useful to analog engineers as precision resistors and capacitors. The process repeatability necessary to fabricate functional corner models for digital transistors across a 300mm or larger wafer necessitate incredibly planar, repeatable processing which helps control the spurious parasitics that normally impact an analog design. One significant caveat is that the higher doped channels, substrates and source/drain areas and the use of source-drain extensions increase the parasitic capacitances and cause a significant reduction in the maximum achievable GBW of analog circuits.

This is relatively easy to observe by inspection given the transition frequency of a common source amplifier [22]:

$$f_H = \frac{1}{2\pi\{R_S[C_{gs} + C_{gd}(1 + |A_{v,LF}|)] + R'_{out}C_{db}\}}$$

where the parasitic capacitances C_{gd} and C_{db} are increased by heavier doping and the parasitic resistance and the source is increased with shallower source and drain junctions.

In summary, digital device engineering solutions to improve submicron digital transistor operation generally have negative impacts on analog circuit operation as scaling continues. Complicating this, specific analog circuits typically have very different process requirements, for example RF circuits require high frequency transistors above all other effects, whereas biomedical applications require low noise and low power process technologies. Since analog only processes exist, but are only economical for specialized processes or products, significant effort has been dedicated to incorporate analog functionality in processes normally engineered for digital design. Among these, the most widely accepted are slight modifications of the process flow to improve analog functionality without affecting the baseline digital process. Of these types of modifications, the inclusion of power transistors has become fairly standard [23].

Even for digital only chips the need for several sets of CMOS technologies on the same die and tighter integration between circuit design are recognized as critical for SoC systems below the 70nm node [24]. The need for low current design is driven by the demand to decrease power consumption. Low voltage operation is required for thinner gate oxides and digital operation and to prevent the devices from operating exclusively in

velocity saturation mode. In fact, contemporary technology limitations to digital MOS scaling is a very active research area [25, 26]. There are several completely new technologies in the research stage [27-29] which differ significantly from MOS transistors, and as such, beyond the scope of this thesis.

Until these technologies are introduced, analog circuit design and migration of current analog circuits from one CMOS process node to the next will continue. Currently, process technology changes migration of analog designs from one technology node to the next requires significant analog design effort and usually requires a complete redesign of the analog portion of the mixed signal circuits. This is one of the reasons cutting edge mixed signal processes typically lag two generations behind the current limit of digital scaling. Optimization of transistor design for digital operation will likely expand this gap. In part this stems from the lack of a systematic design method in circuit design solutions based on the ideal square law behavior of transistors. As process technology scales, the square law behavior of the transistor deviates significantly from ideal. This is one of the reasons that it is common knowledge that analog design takes several fabrication cycles before a product may meet specifications – whereas with digital design the first product is functional.

An Alternative Approach – New Design Methods

Some significant effort has been made to streamline design in submicron analog circuits by revisiting circuit design approaches. Generally, the approach involves

revisiting the fundamental operation of analog circuitry and selecting parameters which affect the fundamental operation of the circuit class for careful consideration. A conservative approach is demonstrated in [16] where an adaptation of the square law behavior in used. This method retains the square law approach to circuit design, but employs the "transconductance efficiency" (log g_m/I_d plotted versus log I_d) figure of merit shown in Figure 2 and lookup tables to predict circuit performance. Using this figure, it is shown that a high transconductance efficiency leads to lower current consumption thus lower power can be achieved by increasing the g_m/I_d ratio while keeping bandwidth (g_m/C_{gs}) constant. This is exactly the point of the "level of inversion" or "g_m/I_d based" design approaches. The key difference is that by maintaining square law behavior (i.e., the constraint of $V_{ds} > V_{gs}-V_t$) to keep devices in strong inversion discards some of the useful voltage swing available to designers, which is critical with falling power supply voltages. This restricted design range (strong inversion, no velocity saturation) is identified in Figure 2 as the "traditional design space" and is equivalent to $V_{gs} - V_t = 2nU_T\sqrt{IC}$. An important point is that transconductance efficiency (g_m / I_d) is the intrinsic transistor gain available for design per unit current. So, while larger gains may be obtained in the strong inversion regime, the gain per unit bias current falls dramatically. This leads the designer towards increasing widths for low power design, not increasing bias current. Since bias current determines the bulk of the power consumption for a circuit, significant power savings can be realized.

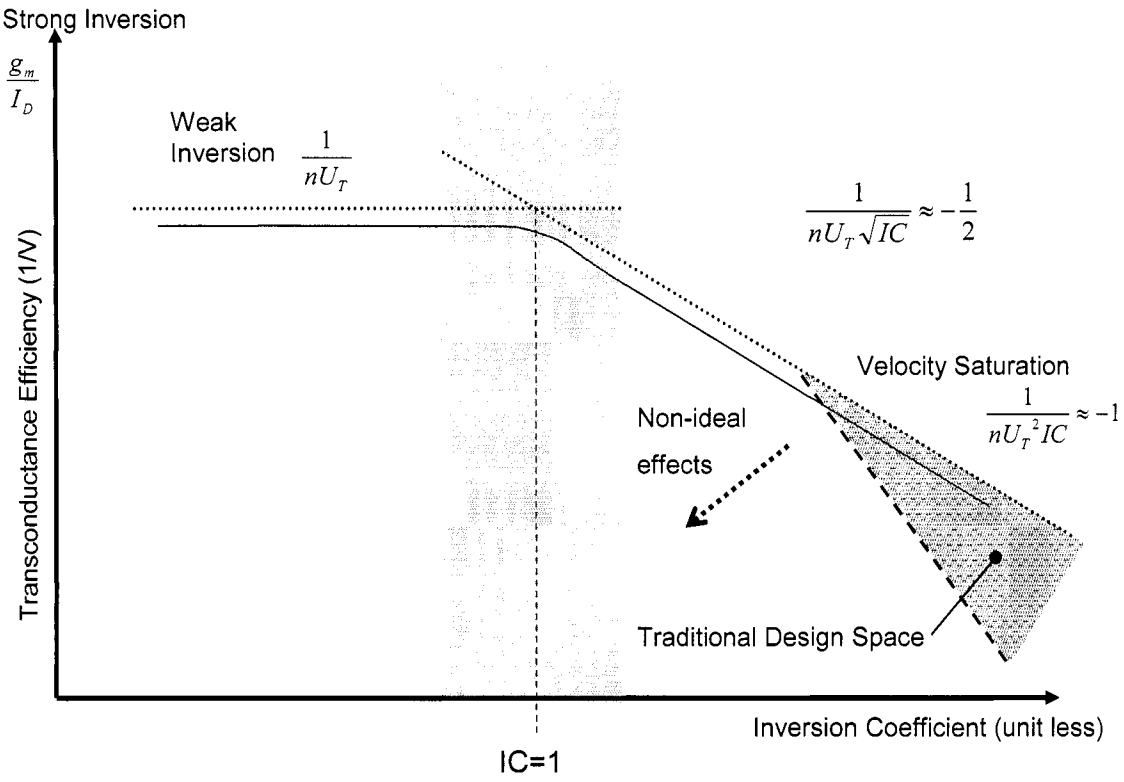


Figure 2. Transconductance Efficiency Curve. IC is Equal to I_d Normalized to Transistor Technology Leakage Current (I_o), Width (w) and Length (l).

The premise of the level of inversion methodology is that designing with the transconductance efficiency ratio will allow collapsing power and frequency requirements to a single variable: *w*, the width of the transistor [30]. Choosing L is based on the circuit design space operation requirements [31]. A simplified L decision tree is shown in Figure 3 based upon the design space variables discussed to this point. With a single set of interrelated relationships directly tied to a simplified transistor model continuous through all levels of inversion, the entire analog design space is captured.

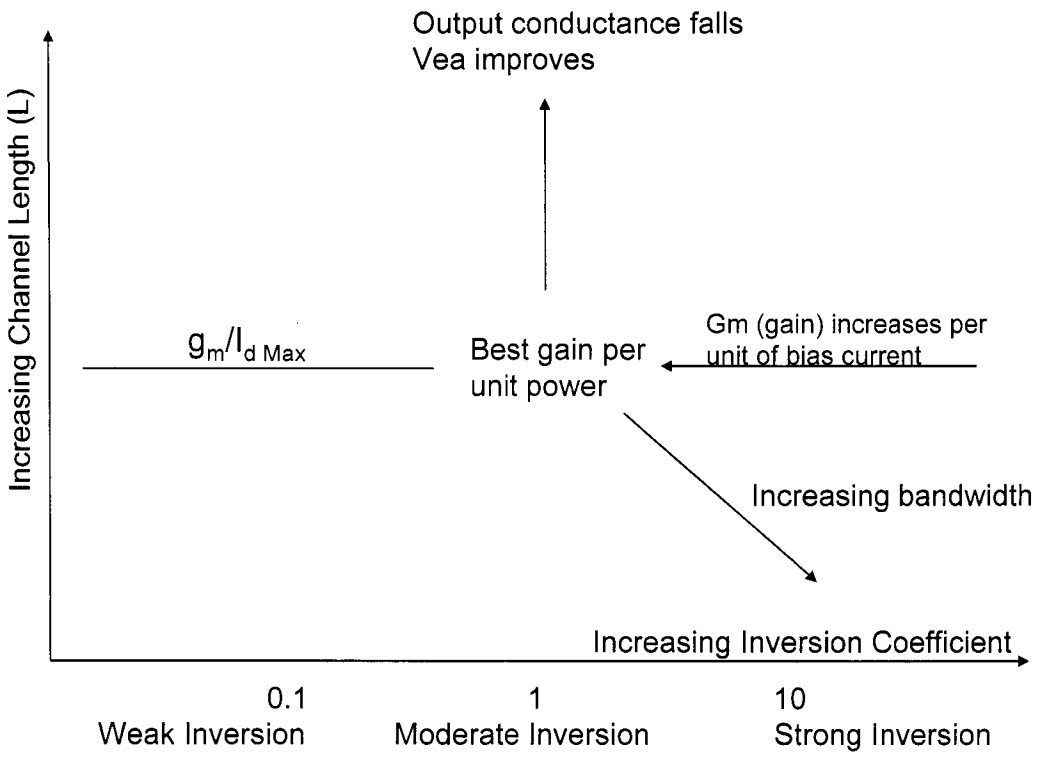


Figure 3. Simplified Selection of Design Length (l) Based Upon Design Goals.

An Illustrated g_m/I_d Design Example

To illustrate this approach an example of g_m/I_d based design is shown below, extracted from the general approach presented in [30]. The pedantic approach is used to provide a specific result and show a specific design case to illustrate the methodology.

Technology parameters for an example process are given as:

$$V_{early} = 8\frac{V}{\mu m}$$
$$U_T = .0259V$$
$$n = 1.35$$
$$\mu_n = 550 \cdot 10^{-4} \frac{V^2}{cm \cdot s}$$
$$\varepsilon_{ox} = 0.345 \cdot 10^{-10} \frac{F}{m}$$
$$t_{ox} = 30 \cdot 10^{-9} m$$
$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} = 0.00115\frac{F}{m}$$
$$K_n = \mu_n C_{ox}$$
$$Io = 2n\mu_n C_{ox} U_T{}^2$$

These equations allow the determination of a specific current, denoted (I_o), characteristic to the process. Basic specifications are then selected for design:

$$f_t = 10Mhz$$
$$|A_v| = 50dB$$
$$C_{Load} = 10pF$$

Length is chosen as a function of the available technology and which operational region is to be optimized (see Figure 3):

$$L_{\min} = L = 2\mu m$$

Next the general expression to relate (g_m/I_d) and the bias current (I_{do}), inversion coefficient (IC) and normalized bias current scaled as a function of transistor size ratio (I') are introduced. These equations would hold for all transistors in a circuit.

$$\frac{g_m}{I_d} = \frac{1}{nU_T} \cdot \frac{(1 - e^{-\sqrt{IC}})}{\sqrt{IC}}$$

$$I_{do} = \frac{g_m}{\frac{g_m}{I_D}}$$

$$I' = \frac{I_{do}}{\frac{w}{l}}$$

$$IC = \frac{I_d}{2n\mu_n C_{ox} \frac{w}{l} U_T{}^2} = \frac{I_d}{I_o \frac{w}{l}} = I' \frac{1}{2n\mu_n C_{ox} U_T{}^2}$$

With these equations the basic design equations for bandwidth (f_t) and gain ($|A_v|$) for a common source NMOS amplifier can be rewritten to emphasize the transconductance efficiency expression (g_m/I_d). As is clearly seen, the gain bandwidth figure of merit ($f_t * |A_v|$) for an amplifier can be expressed in terms of the transconductance efficiency (g_m/I_d), the bias current (I_{do}), the output conductance per unit length (V_{early}) and the load (C_{Load}).

$$f_t = \frac{g_m}{I_D} \cdot \frac{I_{do}}{2\pi C_{Load}}$$

$$|A_v| = \frac{g_m}{I_D} \cdot V_{early} \cdot L$$

This result allows us to determine the transconductance efficiency required from the given specifications:

$$\frac{g_m}{I_d} = \frac{10^{\frac{|Av|}{20}}}{V_{early} \cdot L} = 19.76\frac{1}{V}$$

$$I_{do} = \frac{2\pi f_t \cdot C_{Load}}{\frac{g_m}{I_d}} = 3.177 \cdot 10^{-5} \approx 32uA$$

This allows determination of the inversion coefficient. However, the form:

$$\frac{g_m}{I_D} = \frac{1}{nU_T} \cdot \frac{(1-e^{-\sqrt{IC}})}{\sqrt{IC}}$$

is difficult to use to find a closed form solution for IC. It has been shown that a good approximation is given by [32]:

$$\frac{g_m}{I_d} = \frac{1}{nU_T} \cdot \frac{1}{\frac{1}{2} + \sqrt{\frac{1}{4} + IC}}$$

This is rearranged to give:

$$IC = \frac{nU_T \frac{g_m}{I_d} - 1}{-\left(nU_T \frac{g_m}{I_d}\right)^2}$$

And in this case:

$$IC = 0.6475$$

Which shows the transistor is in moderate inversion, which is generally described as bounded by weak inversion IC<0.1 and strong inversion IC>10 [33].

The current bias required (I'), is determined from the inversion coefficient (IC) and the specific current delivered by a unit sized transistor (I_o) from:

$$I' = I_o \cdot IC$$
$$I' = 7.42 \cdot 10^{-8} \text{ A}$$

Since L was already specified, the last parameter to determine is (W). From the bias current derived to set the transistor in the moderate inversion regime, the gain required and thus the needed transconductance efficiency, a closed form result for (W) results:

$$W = L \cdot \frac{I_{do}}{I'}$$
$$W = 856uM$$

Extensions of the g_m/I_d Method

The preceding example uses a capacitively loaded common source amplifier, but it has been shown that the approach is extensible to any circuit topology [30]. As a note to the reader: [30] also includes design examples for folded and miller compensated operational transconductance amplifiers. Further examples are available in the literature [34, 35]. An application of g_m/I_d based design which incorporates a power minimization strategy is presented in [36]. One interesting result of using the g_m/I_d based approach is that it provides unusual insight into circuit design not usually afforded by simulation driven analysis. In contrast to the "design by sweep" simulation driven methods, novice designers gain direct insight into the operation of the circuit and the fundamental limits imposed by the technology [37, 38]. As a case in point, one of the factors that limit high frequency operation is the parasitic capacitances at the output node. This is shown in

Figure 4 [30]. As the current increases, the width to maintain g_m/I_d eventually results in a parasitic capacitance larger than the load and thus establishes a maximum gain bandwidth frequency. This example also directly shows the power saving available by correct choice of g_m/I_d ratio and bias current, normally obfuscated by design analysis.

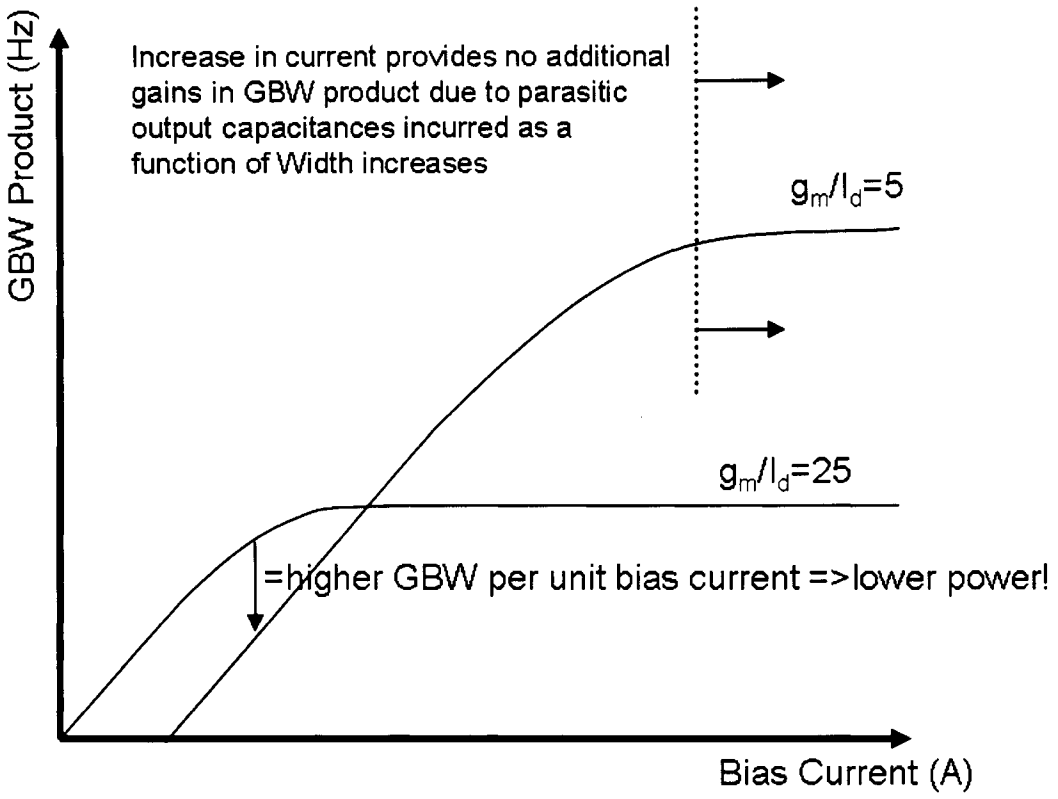


Figure 4. Closed Loop Calculation Lends Insight Into Frequency Limitations.

In [39], the authors use the level of inversion methodology to facilitate the scaling of analog circuits within a process (to reduce power, for example) or across processes. Table 2 summarizes the effects of scaling on both analog and digital circuit operation.

Table 2. Comparison of Digital and Analog Scaling Methods and Effects.

	Digital Scaling Methods				Analog Scaling Methods		
Parameter	Symbol	Constant Field Scaling	Constant Voltage Scaling	Constant Voltage Scaling with Velocity Saturation	Constant Inversion Coefficient Scaling	Constant Length Scaling (strong inversion)	Constant Length Scaling (weak inversion)
Gate length	L	$1/\alpha$	$1/\alpha$	$1/\alpha$	1	$1/K_L$ $(1/\alpha)$	Same as C.L.S.I.
Gate Width	W	$1/\alpha$	$1/\alpha$	$1/\alpha$	K_v^2/K_{ox} (α)	$K_v^2 K_L/K_{ox}$ (α^2)	Same as C.L.S.I.
Field	E	1	α	α			
Oxide Thickness	t_{ox}	$1/\alpha$	$1/\alpha$	$1/\alpha$			
Substrate Doping	N_α	α^2	α^2	α^2			
Gate Capacitance	C_g	$1/\alpha$	$1/\alpha$	$1/\alpha$			
Oxide Capacitance	C_{ox}	α	α	α	K_{ox} (α)	K_{ox} (α)	Same as C.L.S.I.
Transit Time	T_r	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha$			
Transit Frequency	Fr	α	α^2	α			
Voltage	V	$1/\alpha$	1	1	$1/K_v$ $(1/\alpha)$	$1/K_v$ $(1/\alpha)$	$1/K_v$ $(1/\alpha)$
Current	I	$1/\alpha^2$	α	1	K_v^2 (α^2)	K_v^2 (α^2)	K_v^2/K_L^2 (1)
Power	P	$1/\alpha^2$	α	1	(α)	(α)	$(1/\alpha)$
Pdp	PDt	$1/\alpha^3$	$1/\alpha$	$1/\alpha$			

The (α) symbol in the analog scaling portion of the table results when all K scaling factors are equivalent. Decoupling the K scaling factor allows scaling within a process to reduce voltage or power consumption. It is observed that fixing a single variable, such as electric field, or inversion coefficient, and then scaling to maintain that fixed variable constant has many impacts upon both the design of the transistor and the performance benchmarks of the transistor.

Since the g_m/I_d method is closed form, there are labs developing computer assisted design tools to assist the design process and facilitate insight [37, 40, 41]. Some classes of circuits can even be synthesized all the way to layout [41]. The tools mentioned here are a small subset among many reported in the literature for automation of analog design or layout, but these three tools mentioned specifically differ significantly from the others in that they function primarily to lend insight on the operation of the transistor in the selected design space to the designer, embrace a continuous model for weak through strong inversion and are available free via the world wide web.

Two important caveats should be mentioned in employing this design approach. The analytical derivation of the circuit transfer function for the parameters of interest is a key requirement. The circuit design equations must be solved in terms of g_m to be able to properly correlate the inversion coefficient to the design goals. Commercially available tools which perform symbolic analog analysis will greatly aid this process, examples are SYMBA or ISAAC and Analog Insydes [42, 43]. Secondly, designers must be aware that in weak inversion, and generally in this approach, it is necessary to employ current biasing rather than voltage biasing the transistor gates. Since generating the correct level of inversion is important for proper function of these circuits and process drift typically results in some variation of the intrinsic specific current, circuits which generate a current bias as a function of specific current are recommended. A nice example by Linares-Barranco is presented in [44]. Current biasing is counter to industry standard and is in process of being verified in David Parent's research lab.

Modeling Continuously Through the Inversion Region

Revisiting circuit design methods highlights the need for a MOS transistor model which can continuously predict behavior from strong to weak inversion and account for higher order submicron physical effects. The frustration in the analog community with higher order effects in low dimension technology nodes is pervasive. Indeed, Razavi opines that scaling accelerates model inaccuracy [11] and suggests complete arrays of analog circuits for characterization purposes rather than relying on modeling. Partly, this frustration is driven by the successive empirical modifications of square law models to correct for submicron effects and the prevalence of the Berkeley short-channel insulated gate field effect transistor model (BSIM) parametric models which model the weak inversion saturated operation and the strong inversion saturated operation of the MOS transistor discontinuously. The current version of BSIM transistor models has some 400 model parameters, requires a week of time from an experienced parameter extraction engineer and may also include several bins of parameters within a single model. Indeed, it has been observed:

> *The BSIM models are a never-ending string of approximations that appear to accomplish more the need for generating PhD candidates in ECE at UC Berkeley than accomplishing good device modeling techniques [45].*

Generally the reason engineers and scientists model phenomenon is to simplify to phenomena they are trying to study or to predict the behavior of a system. In science,

modeling generally lends insight. In engineering, modeling is used to accurately predict the behavior of the system which is being designed. There are subtle but significant differences between the two. In either case if a system is not completely understood then models based upon experimental results can be employed. Empirical models are derived from experimental data – using additional parameters to "fit" the model behavior to the observed data. The case of MOS transistors is an excellent example. The first models for the simulation program with integrated circuits emphasis (SPICE) were derived from physical models. These models greatly simplified the "simulation" of the final circuit without tedious hand calculations and were based upon physical understanding of the transistor. However, since the model was simplified, the final circuit did not always match the simulation. In order to combat this, successive approximations based upon empirical evidence were incorporated until the models became fairly cumbersome. This trend is captured in Figure 5, modified from [46]. Compact models, on the other hand, are constantly under development toward four primary goals; to accurately predict the behavior of the transistor before fabrication, reduce the amount of data which needs to be gathered to create an empirical model, improve understanding of the transistor, and reduce the computational demands of transistor level simulation.

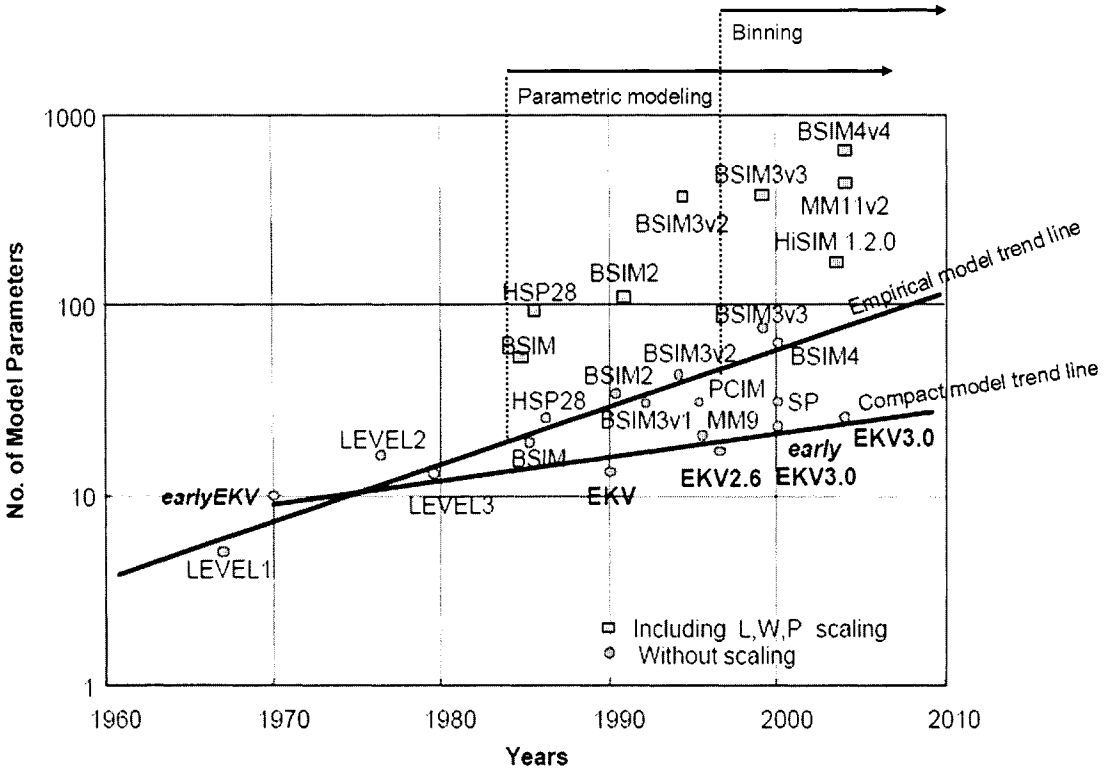


Figure 5. Parametric Model Trends. Modified from [46].

The EKV Model

One of the most prevalent examples of compact models is the Enz Krummenacher Vittoz (EKV) model [47]. The EKV model was derived with low power and low voltage operation of the MOS transistor in mind and oriented towards facilitating analog design. Many models use the source as a reference node, but the EKV model uses the bulk as a reference node. This allows modeling of the inversion charge relationships rather than using a source tied charge sheet approach and treats the MOS device as a symmetrical device. Thus the device can be simulated symmetrically, where the source and drain can have separate voltages and the transistor may operate in either direction. Because the

inversion region in the channel is modeled as a sheet charge with respect to the bulk, current flow and channel potential are modeled continuously from the weak to strong inversion and from linear to saturation. Surface potential based modeling is another approach [48], but it is suggested this is a special case of the EKV model [49]. These approaches differ in their derivation and as such, the formulation for IC differs slightly. The higher level design approaches, however, are generally cross applicable.

In the analysis in this thesis the EKV model is highlighted for several reasons. First and foremost, it is a physically based model – understanding the model leads to understanding of the transistor operation and subsequently the circuit operation. The approach used in developing the model leads to the circuit design approach reviewed above so evaluation of the device operation and generation of the figures of merit for analog design are intimately connected. The EKV model has a small number of parameters and thus is computationally efficient. While this would impact larger circuits, it also eases both comprehension and extraction. Most importantly, the single continuous interpolated analytical expression for current from strong to weak inversion is very accurate and models sub threshold behavior more accurately than BSIM models. This in turn leads to very accurate prediction of moderate inversion behavior [50].

One of the features that facilitates EKV model adoption is that the model implementations in SPICE analysis packages are widely available, the model is well supported and also well documented While other models which take a similar approach are mentioned in the literature, notably the SP Model [51] and the ACM model [32] these models do not appear to be as well adopted as the EKV model. In particular, the SP

model may predict MOS behavior more accurately, but extraction routines and simulator support is not as robust. Besides being supported under the bulk of SPICE circuit simulators, there are publicly available behavioral models and device extraction packages available from Silvaco (UTMOST), Agilent (ICCAP-Admos) and Synopsis (Aurora). Extraction of parameters is outlined in [52] with a short example in [53]. More importantly there are tools under development to allow conversion from more popular BSIM models without re-extraction of model parameters [54]. This is critical because analog design is most often done in multi-user foundries and an accurate and convenient way of converting industry standard BSIM models or support at the foundry for advanced model extraction is critical. Currently, the EKV model has both of these features. In addition, the model is under continuous and active development, with the latest version for modeling UDSM effects released as version 3. Finally, there is extensive support for behavioral modeling in C, VHDL-AMS and Verilog-A.

One of the most important features of the EKV model is the definition of a "pinch off voltage." V_p is defined as the transition voltage, and can be correlated to the V_t by:

$$V_p \cong \frac{V_g - V_{to}}{n}$$

Since (n) is a function of the process variables γ and φ, the pinch of voltage is a then function of V_g and V_{to}. This allows the unambiguous extraction of γ and φ. On one side of V_p the inversion charge exists with nearly constant density (i.e. strong inversion). One the other side the inversion charge disappears exponentially with the changing voltage (i.e. weak inversion). V_p is used to describe both the subthreshold behavior and current

saturation condition. This simplifies the discussion of the regions of operation. When both V_s and V_d are below V_p, a strong inversion region exists in the entire channel. In the EKV model this is referred to as "conduction" which is more commonly identified as the linear region of operation. When V_d exceeds V_p or V_s is less than V_p then the channel pinches off and the device operates in saturation mode. The only difference is the direction of current flow – again demonstrating the symmetry of the model. Current flow direction is described as the sign of (V_d-V_s). When V_d and V_s are both larger than V_p then the entire channel is pinched off – so no conduction occurs. Subthreshold conduction occurs if either V_d or V_s is slightly larger than V_p.

If either V_d or V_s exceeds the junction potential then the junction is forward biased. This activates the bipolar mode in the device. Readers interested in the bipolar operation of the MOS transistors are referred to [55-60]. Bipolar operation would have to be modeled independently, as MOS models do not accurately model the parasitic bipolar operation. These devices are sometimes employed in ultra low noise feedback topologies [61], or CMOS process compatible devices for bandgap circuits, although CMOS only bandgap circuits have been shown to be a competitive technology, especially for sub-volt designs [62]. As an aside, there are indeed BJTs available in the Spartan CMOS process, although the configuration varies. Another useful feature of the EKV model is the inclusion of "matching parameters" which are critical for analog design to facilitate matching analysis. This greatly reduces the computational demand during Monte Carlo simulation via the parameters AVTO, AGAMMA and AKP. There is no need to create separate parameters sets for each geometry set.

With a reduced parameter set come additional benefits. Since the extraction procedure is simplified and links directly and unambiguously to the design flow, extracting a core set of parameters can be useful for gaining insight into the function of the transistor and its performance in the circuit design space. The figure of merit for this design approach is the transconductance efficiency curve (Figure 2) which is very much a function of the intrinsic transistor properties. By identifying the level of inversion present in the MOS transistor and extending analysis from this common point, ALL transistors across ALL processes may be compared. By evaluating the change in transconductance efficiency as a function of the inversion coefficient insight into the transistor operation is gained and links directly with circuit design.

Comparing Technologies Using Level of Inversion Analysis

In using the G_m/I_d method to compare technologies, it becomes clear how to improve the performance of analog transistors through device engineering. Since the simulation requires less extensive analysis than generating a complete BSIM model and using the model to generate test designs, prediction of analog performance in a general class of solutions (i.e. common source amplifier performance) then evaluating those predictions through simulation greatly streamlines the transistor engineering device design flow.

Unfortunately, due to the unacceptable level of risk involved, the process engineering community is very slow to adopt changes in the baseline process, especially for analog design, as the effective chip area of the analog portion of design is rather small. One notable exception is the need for power transistors. As system integration continues there

are many applications for voltages that exceed the specification for standard digital transistors. These include EEPROM and drivers for off chip transducers such as small motors, or microphones and speakers for audio applications [23]. MEMS systems often require high voltages to produce the electrostatic fields necessary for actuation of small, relatively stiff microstructures [63]. Both piezoelectric and ultrasonic transducers require large actuation voltages and thermal inject printers and nitonol applications both require relatively large driver wattage. TFT Drivers often have requirements for switching tens of volts [64]. Finally, it is important to note that with shrinking oxide thickness and shorter channels, even 5V applications such as I/O may exceed the operating range of standard digital transistors. An excellent review of power devices in CMOS compatible processes is in [23].

High Voltage, High Current Transistor Design

Since the design of power transistors is driven by high voltage applications, the next section will briefly review breakdown effects in MOS transistors and discuss the design strategies for mediating these effects. Aspects of high voltage design will be shown to have direct implications for low voltage and low current applications.

Junction Breakdown

Since a transistor is made up of pn junctions the breakdown of these junctions is critical to understanding transistor breakdown. Two major effects drive junction breakdown; avalanche and zener breakdown.

Avalanche Junction Breakdown

Avalanche breakdown occurs when the electric field is large enough to accelerate the carriers to the point that they gain more energy in the electric field than they lose through lattice impacts. Avalanche breakdown is a function of the lighter doped side of the junction. The critical electric field is calculated by [65]:

$$\varepsilon_{critical} = \frac{4E5}{1 - \frac{1}{3}\log\frac{N}{1E16}}$$

$$I_r = M * I_0$$

$$M = \frac{1}{1 - \left(\frac{V_r}{BV}\right)^n}$$

$$BV = \frac{\varepsilon_s (N_a + N_d)}{2qN_aN_d}\varepsilon_{crit}{}^2$$

Where N is the doping concentration on the more lowly doped side of the junction, V_r is the reverse bias voltage, and n is experimentally determined (about 4 for n+/p and 6 for p+/n junctions). From the above equations, as V_r approaches breakdown voltage (BV), the multiplication factor (M) becomes very large. Thus, heavier doping in the lowly doped side results in a lower breakdown voltage. Shur gives an alternate estimate based on fields [66]. Creating dually heavily doped junctions is not a direct solution because it leads to another form of uncontrolled current conduction, discussed below. Avalanche diodes can be engineered to have breakdown voltages up to several thousands of volts.

Zener Breakdown

When both sides of a junction are heavily doped, the total depletion width does not appreciably change with application of a reverse bias. In addition, the conduction band of the p-side and the valence band of the n-side are very near in potential at equilibrium. In this case application of even a small voltage may cause electrons to tunnel through the junction. This is essentially a quantum level effect in which the location an electron exists on either side of the junction is expressed as a probability. Streetman [67] employs the simple covalent bonding model in which the ionization of the electrons in the covalent bonds of the host atoms on the p side of the junction are considered. At a high enough field strength (10^6 V/cm) electrons are ionized and accelerated toward the n-side of the junction generating a reverse bias current [67]. Typically this occurs when junctions are doped at 10^{18} or above on both sides and the breakdown voltage is typically below 5.6 Volts. Some confusion exists in the field describing this effect. The terms "Zener diode" and "Breakdown diode" are often used interchangeably even though they represent different physical operation. In fact, both effects are present in diode junctions. Most commercially sold "Zener" diodes are actually avalanche breakdown diodes.

Breakdown in MOS

Punch Through

Punch through occurs when the drain and source depletion regions in the channel contact. Due to surface effects on the electric field, this actually occurs below the surface

of the transistor (hence the inclusion of the ψ term). A rough estimate may be obtained by adding the depletion widths. A closer estimate may be obtained from the following formulae [68]:

$$V_{PT} = \frac{qN_a}{2K_s\varepsilon_0}\left[L_{eff} - \sqrt{\frac{qN_a}{2K_s\varepsilon_0}(V_{sb} + V_{bi})]}\right]^2 - (V_{sb} + V_{bi})$$

$$L_{eff} = L_{actual} - y_{source} - y_{drain}$$

$$y_s = \left[\sqrt{\frac{2K_s\varepsilon_0}{qN_a}(V_{bi} - \psi_s)]}\right]$$

$$y_s = \left[\sqrt{\frac{2K_s\varepsilon_0}{qN_a}(V_{bi} - \psi_s + V_d)]}\right]$$

$$V_{bi} = \frac{kT}{q}\ln\frac{N_aN_d}{n_i^2}$$

$$\psi_s = 2\frac{kT}{q}\ln\frac{N_a}{n_i}$$

It is important not confuse the effective length and the drawn length. The drawn length does not include the S/D undergate diffusion, so L_{actual} is typically the "true" distance of the source diffusion to drain diffusion channel length. Unfortunately, L, L_{eff}, L_{actual}, L_{drawn} are used interchangeably. Shur [66] has a nice alternative approach based upon currents. In short channel devices the surface and drain potentials may exceed the channel length. In this case, drain current is dominated by the space charge limited current, and punch through can occur in both linear and saturation regions. The following formulas provide an estimate of this current in both regions as a function of drain voltage:

$$I_{sp\,lin} = \frac{9\varepsilon_s \mu_n A_{s/d} V_d^{\,2}}{8L^3}$$

$$I_{sp\,sat} = \frac{2\varepsilon_s v_s A_{s/d} V_d}{L^2}$$

where v_s is the carrier saturation velocity. From the above equation, it is clear that increasing the channel doping would reduce punch through effects. This is the approach used in power MOS devices such as the double diffused MOS transistor (DMOS). However, this increases the parasitic capacitances. As a result, in modern devices a substrate doping region to decrease the effects of substrate punch through is often employed [69].

Hot Electron Effects

Hot electrons are actually of two main classes. Substrate hot electrons are thermally generated carriers which gain energy in the vertical electric field. Channel hot electrons are generated at the surface in the high electric field near the drain. At high electric fields (MOS operated in saturation, high V_{ds}) the channel is pinched off. The large electric field near the drain results in the generation of hot carriers. It must be noted that hot holes are generated at higher voltages than hot electrons, so a PMOS would be more tolerant to large V_{ds}. Both classes of hot electrons contribute to hot electron effects.

Hot electrons have essentially four main effects. They may become embedded in the oxide, shifting the threshold voltage. They may jump the 3.2eV oxide barrier and generate a gate current. They may impact the lattice in the Si-$Si0_2$ interface damaging it and causing interface charges, shifting the threshold voltage and leading to increased

surface state traps resulting in a higher noise figure. They may generate a substrate current and overload the bias generator, forward biasing the source substrate junction and initiating an avalanche breakdown event.

Obviously, reducing the electric field in the vicinity of the drain will reduce hot electron effects. Lightly Doped Drain (LDD) structures are one way to accomplish this. As an interesting note, LDD transistors fell out of favor in the 90's with the revisitation of constant field scaling at about the 0.35um node [70].

Avalanche Breakdown

Avalanche breakdown involves the generation of generation of hot electrons [71]. This is illustrated by assuming the gate and back gate are connected to source. At a high enough voltage, the drain source voltage breaks down. Avalanche multiplication injects large numbers of majority carriers into the lightly doped back gate, which results in the debiasing of the back gate. As soon as the back gate debiases the source back gate junction injects minority carriers into the back gate. These minority carriers flow to the drain, causing further avalanche breakdown. Streetman [67] states that the lightly doped region determines the avalanche breakdown region. If it is short in relationship to the minority carrier diffusion length, then avalanche effects will be increased.

Oxide Breakdown and Leakage

A gate oxide will break down 6-7 * 10^6 V/cm. (about 6-7 V on a gate of 100 Angstroms) [67]. This tends to be cumulative effect. Much of this has to do with the

"quality" of the gate oxide. A poorly formed gate oxide will have more free bonds and be more likely to conduct current when an over voltage is applied. In addition, very thin gate oxides may allow direct tunneling of electrons in a fashion similar to Zener breakdown. This is generally referred to as gate induced drain leakage (GIDL).

Power Transistor Design

The invention of the power metal oxide semiconductor field effect transistors (MOSFETs) was partly driven by the limitations of power bipolar junction transistors (BJTs), which, until recently, was the device of choice in power electronics applications. The bipolar power transistor is a current controlled device. A large base drive current as high as one-fifth of the collector current is required to keep the device in the ON state. Also, higher reverse base drive currents are required to obtain fast turn-off. Despite the very advanced state of manufacturability and lower costs of BJTs, these limitations have made the base drive circuit design more complicated and hence more expensive than the power MOSFET. Finally, integration of bipolar devices is an expensive extension to baseline MOS processes, whereas with minor modifications MOS devices can be engineered to have power transistor characteristics.

Another BJT limitation is that both electrons and holes contribute to conduction. Presence of holes with their higher carrier lifetime causes the switching speed to be several orders of magnitude slower than for a power MOSFET of similar size and voltage rating. BJTs may also suffer from thermal runaway. This occurs when the forward

voltage drop decreases with increasing temperature-causing diversion of current to a single device when several devices are paralleled. Power MOSFETs, on the other hand, are majority carrier devices with no minority carrier injection. They are superior to the BJTs in high frequency applications where switching power losses are important. Plus, they can withstand simultaneous application of high current and voltage without undergoing destructive failure due to second breakdown. Power MOSFETs can also be paralleled easily because the forward voltage drop increases with increasing temperature, ensuring an even distribution of current among all components. However, at high breakdown voltages (>200V) the on-state voltage drop of the power MOSFET becomes higher than that of a similar size bipolar device with similar voltage rating. This makes it more attractive to use the bipolar power transistor at the expense of worse high frequency performance. Over time, new materials, structures and processing techniques are expected to raise these limits.

The Structure of a MOSFET

In the lateral channel design, the drain, gate, and source terminal are placed on the surface of a silicon wafer. This is suitable for integration but not for obtaining high power ratings since the distance between source and drain must be large to obtain better voltage blocking capability. To appreciate this fact, recall that the drain current of an n-channel MOSFET operating in the saturation region is given by:

$$i_D = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS} - V_t)^2$$

It follows that to increase the current capability of the MOSFET its width, W, should be made large and its channel length, L, should be made as small as possible. Unfortunately, reducing the channel length of the standard MOSFET structure results in a drastic reduction of its breakdown voltage. Specifically, the depletion region of the reverse-biased body-to-drain junction spreads into the short channel, resulting in breakdown at a relatively low voltage. Thus the resulting device would not be capable of handling high voltages typical of power-transistor applications. For this reason, new structures had to be found for fabricating short channel MOSFETs with high breakdown voltages.

Also, the drain-to-source current is inversely proportional to the length. With the vertical channel design, the drain and source are placed on the opposite sides of a wafer. This is suitable for a power device, as more space can be used as source. As the length between the source and drain is reduced, it is possible to increase the drain-to-source current rating, and also increase the voltage blocking capability by growing the epitaxial layer (drain drift region). The vertical MOSFET (VMOSFET) design was the first to be commercialized having a V-groove at the gate region. The diffused MOSFET (DMOSFETs) replaced VMOSFETs as there were stability problems in manufacturing, and they also had a high electric field at the tip of the V-groove. The DMOSFET design has a double-diffusion structure with a P-base region and an N+ source region. The u-groove MOSFET (UMOSFET) design is similar to the VMOSFET in this design has a U-shaped groove at the gate region. It has a higher channel density, which reduces the on-

resistance as compared to the VMOSFETs and the DMOSFET. UMOSFET designs with the trench etching process were commercialized in the 90's.

Double Diffused MOS (DMOS)

At present, the most popular structure for a power MOSFET is the double diffused (DMOS) transistor [72-74]. The device can be fabricated vertically or laterally, but lateral implementations are easily incorporated into mixed signal CMOS processes. The general operation of the DMOS is as follows: application of a positive gate voltage, $V_{gs} > V_t$ the threshold voltage induces a lateral n channel in the p-type region underneath the gate oxide. The resulting channel is short. Current is then conducted from the source through the resulting short channel into a lightly doped conduction region and then to the drain.

Despite the fact the DMOS transistor has a short channel; its breakdown voltage can be very high (as high as 600 V). This is because the depletion region between the p region and the body extends mostly in the lightly doped n region and does not spread into the channel. The result is a MOS transistor that simultaneously has a high current capability (50A is possible) as well as the high breakdown voltage just mentioned [75]. Typical V_{ds} breakdown voltages in submicron processes range from 80 – 200V with 20V devices possible in UDSM processes.

In the DMOS (double-diffused MOS) structure the channel length L is determined by the higher rate of diffusion of the p dopant (e.g., boron), compared to the n+ dopant (e.g., phosphorus) of the source. The channel is followed by a lightly doped drift region.

Another version of DMOS is made by implantation. Double-implanted MOSFET (DiMOSFET) forms its source and drain by using a polysilicon gate as mask. The gate is tapered and the p+ shield region is shaped by implantation through the tapered gate. The DiMOS structure improves the control in DMOS structures.

The DMOS structures can have very short channels and do not depend on a lithographic mask to determine channel length. Both structures have good punch through control because of the heavily doped p-shield. The V_{ds} results in a channel depletion region that now lies to a larger extent in the lightly doped region. The lightly doped drift region minimizes the voltage drop across the region by maintaining a uniform field (10^4 V/cm) to achieve velocity saturation. The field (y-component of electric field) near the drain is the same as in the drift region, so avalanche breakdown, multiplication, and oxide charging are reduced. This lightly doped region counters the problems of higher substrate doping levels, shallow junction depths, and high electric fields, which cause breakdown problems with the drain-substrate junction. The drift region increases the punch through voltage, because the depletion region extends into the drift region, rather than into the channel region [76].

Compared to conventional MOSFETs, the DMOS has a higher breakdown voltage, and the substrate current is reduced by a factor of 30. The peak field (x-component of electric field) is shifted away from the oxide so that hot-electron injection into the oxide is much less. The lightly doped drain means a thinner depletion layer, thus less charge sharing and less threshold reduction in short-channel devices. There is also less overlap capacitance, resulting in a faster circuit.

The advantages of the DMOS structure do not come without a price. The threshold voltage, V_t, is more difficult to control in DMOS. Threshold voltage, V_t, is determined by the maximum doping concentration, NA_{max}, along the semiconductor surface. Since the DMOS is not a self aligned structure, and thermal budget determines the channel width and mask to mask alignment determines the initial junction location, variation in channel doping can be problematic. The most critical of these impacts is that a varying NA_{max} leads to variations in V_t. The localization of punch through control to a thin shield region requires a higher doping level, which leads to poorer turn-off behavior for DMOS. The lowly doped drift region increases the series resistance, so the available drain current is reduced. In an addition, the higher voltage means higher power dissipation. For this reason, one must consider the power-delay tradeoff as well as heat generation of a particular design [77].

Laterally Asymmetric Devices

The double diffused devices are actually the most common type of a class of devices termed "laterally asymmetric devices" (LAC). These are transistors with a doping gradient along the length of the channel, usually higher at the source end. Originally conceived as a way of making channels narrower than contemporary photolithographic limits, they eventually fell out of favor for digital design when lithographic techniques improved because of the lack of control over the threshold voltage due to thermal process variation [78]. However, because of the asymmetric structure of the channel, superior breakdown voltage performance is observed. With the shrinking channel length

dimensions transistor designers have begun to create laterally asymmetric transistors because their unique structure offers some advantages in analog performance. LAC devices may be considered as two transistors in series with a continuously increasing V_t, highest at the source. Thus, the highly doped transistor (HD) at the source end of the transistor can be considered the device which modulates the channel current, while the lowly doped transistor at the drain end and can be considered as a variable conductance with a gain of g_m/g_{ds}. Early voltage estimation would then be predicted as [79]:

$$Vea_{total} = Vea_{HD} * \left(\frac{g_m}{g_{ds}} \right)_{LD}.$$

Further benefits are gained from the drift region which has the effect of reducing the velocity saturation by reducing the electric field, as the depletion region extends into the lowly doped side of the junction, rather than the channel region. This also decreases the output conductance. A similar effect was shown to benefit analog transistor design from increased channel doping [80].

Methods

Overview

With increasing system demands on the high density, multi-functional integrated circuits (IC), the need to integrate distinct functions (i.e. digital, analog, high voltage, etc.) on a single chip has become a technological driver. By combining power transistors with the high-performance low-voltage CMOS process, this integration approach provides an excellent method for achieving high performance, mixed power functions on the same chip. Since these devices are already incorporated into the CMOS process flow and should display superior analog characteristics compared to a standard CMOS transistor design, a simulation, fabrication and test cycle was implemented to explore this case.

The procedure described here represents an exploration of the DMOS power transistor that is simulated, fabricated, and tested in a CMOS compatible process. Using Cadence Design Suite to create a GDS2 layout and mask files, the LDMOS was made to approximate the AMIS 1.6μm standard process. Silvaco Maskviews was used to take section or cut-lines through various meridians of the mask files. The "cutlines" were imported into Silvaco Athena software, which was utilized to create a generic rundeck, which simulated the process parameters for the fabricated device. The simulated data was checked against data collected in the lab, and the model was refined to match actual process data. Following this the Silvaco Atlas program was used to perform simple electrical device simulation. The devices were fabricated at Spartan Semiconductor

during the EE167 course fabrication cycle, and then tested using a HP4156A semiconductor parameter analyzer. Testing of the devices highlighted yielded results consistent with prediction.

Design Approach

From the g_m/I_d analysis example, it is clear that reducing the parameter, n, and increasing the technology leakage current, I_o, will improve the low power intrinsic gain of a MOSFET. This is reflected in the g_m/I_d curve as a shift of the curve up with a decreasing n and a shift to the right for increasing I_o, and thus a decreasing the equivalent required width. This is an important point because, as is shown in the detailed example, the width for transistors in moderate inversion and below can become quite large. Another way to consider this is that decreasing n increases the absolute gain per unit width, while increasing I_o increases the current available in moderate inversion to drive a load. Since n is approximated by:

$$n = 1 + \frac{C_{depl} + C_{it}}{C_{ox}}$$

and C_{it} is negligible [81]; reducing the depletion region capacitance is one method to improve gain. I_o is also a function of substrate doping according to the effect of doping on mobility, μ, thus it would seem logical to reduce the substrate doping. This is clearer using the formulation from [4]:

$$n = 1 + \frac{1}{2C_{ox}}\sqrt{\frac{q\varepsilon_e N_b}{\phi_f}}$$

The situation is slightly more complex than this, as presented in [82]. Carriers deep in the substrate increase the effective distance that the gate electric field extends into the substrate, decreasing C_{ox} and therefore decreasing n. However, the carriers in the channel must be under the control of the gate. This is in part captured by the inclusion of n in the I_o derivation. Vittoz, one of the authors of the EKV model, presents a unique way of quantifying this effect in [83], where they employ a C_g vs. V_g curve to measure n.

Reducing the channel, or bulk doping, to an optimum would seem a solid transistor design approach. However, reducing the channel doping has very negative impacts on the early voltage, as channel length modulation effects are increased as the doping in the channel falls. From the analysis presented in the introduction, a significant decrease in the output conductance caused by an increase in channel length modulation effects has negative effects on the intrinsic gain of the transistor. This would impact any amplifier design by lowering the overall gain of the circuit.

Power transistors are commonly incorporated in small dimension processes, so the results are both practical and reasonably available for incorporation into circuit design. Since power transistors mediate channel length modulation effects by incorporation of drift regions and/or lateral source drain doping concentration it seemed logical to investigate their low power operation regime. Given that the power transistors have a high output conductance, a lower channel doping than bulk transistors in the same process, but have not been characterized in the low power regime, opportunities for using the unique structure of power transistors for low power operation with high gain should be explored.

Referring to the g_m/I_d analysis presented it is clear that improvement in the early voltage (V_{ea}), or reduction of the output conductance would increase the gain of the common source amplifier and other amplifiers which use the drain and source terminals as output nodes. LDMOS power devices typically have a very low output conductance. In fact, increasing the output conductance so these devices perform as more ideal switches, as in the digital regime, is a routine design goal for LDMOS devices. As early voltage is evaluated by [35]:

$$V_{ea} = \frac{dI_{ds}}{dg_{ds}}$$

it is clear that reducing the effect of the drain voltage on the channel will result in a larger output conductance. One approach is arrays of transistors where arrays of transistors are used to reduce the electric field per unit width at the drain end [84] lowering the output conductance. It is shown in [79] that devices with an asymmetrically doped channel provide a single transistor solution. By comparison, power transistors are designed with an inherently low output conductance to reduce thermal effects and increase switching speed. LDMOS devices serve as a particular case of asymmetrically doped transistors in which low output conductance for power dissipation reasons coincides with low power, high gain design goals.

However, LDMOS devices have not been widely investigated for low voltage applications. In low power design, the subthreshold current and low output conductance can actually be of use. Since maximum gain is available from transistors operating in the weak inversion region of saturation, high gain power efficient circuits are available in

subthreshold. This effect is facilitated by the doping profile of the transistor, going from high doping at the source to low doping at the drain. Since the subthreshold devices are diffusion devices, a prediction that the doping profile enhancement would behave in a similar manner to BJTs is reasonable. The buried layer length changes as a function of the depth of the inversion layer, which acts as a retrograde well doping. Because of the drift region, this effect differs from standard MOS.

Discussion of the MOS Transistor as a Bipolar Device

One of the important features of these devices to note is the similarity to laterally diffused bipolar junction transistors. In fact, the MOS had been explored as a gate controlled lateral BJT first in [85]. A trench isolated device very similar to a lateral DMOS structure with a direct ohmic contact to the channel region is presented in [86]. In either case, these devices are suggested as they have roughly three orders of magnitude less input referred noise than MOS device. This comes at the cost of a decrease in input impedance and power consumption. Most importantly, it is common to use dopants as lifetime killers in modern submicron processes to reduce latchup due to BJT coupling in the substrate. This would greatly reduce the acivable gain and thus the operation of a MOS device as a BJT.

Simulation

The intent of the process simulation was to accurately model the devices, which had been fabricated using the Spartan Semiconductor Pwell CMOS Process. There were

actually two separate mask sets. The first was from a prior project and the second was a set of masks altered to include new design rules and changes to improve operation predicted by simulation. Since this device was fabricated in a teaching lab and processing is not performed continuously and there is some variation inherent in the process, data was carefully collected during the semester in an effort to account for this variation in the simulation. Some effort to make the process robust during manufacture (as in Design for Manufacturability) was investigated and employed where possible.

The majority of the effort in simulation was in gaining familiarity with the tools available and understanding their appropriate implementation and limitations. Cadence design suite (CDS) is a widely used commercial package, which has excellent capabilities for bottom-up circuit design. There is, unfortunately, no process simulation package available as part of this suite. Since it is widely used in industry, as well as here at San Jose State University, it was important to understand the interface between the process simulation packages and the CDS. The interface is reasonably generic and while specific issues will be mentioned below, the overall approach should hold.

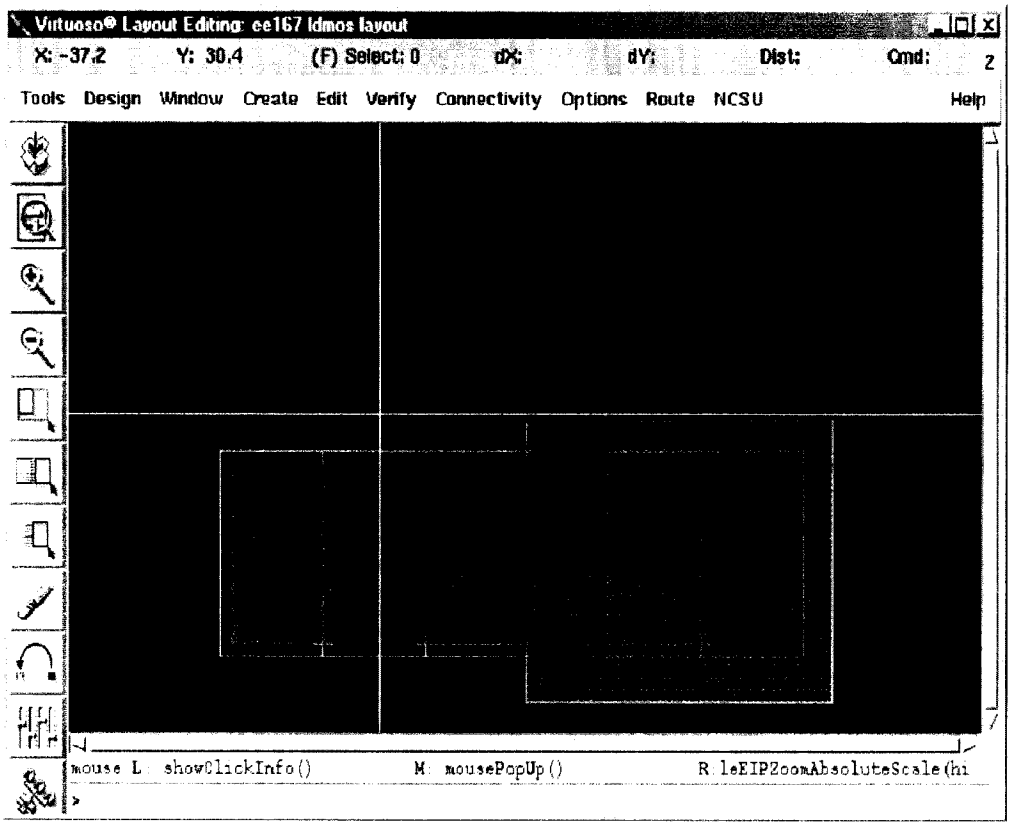


Figure 6. LDMOS Layout from Cadence Virtuoso Layout - Rev. 1

For the CMOS compatible LDMOS, the Cadence layout was suggested by Dr. David Parent and subsequently modified as a function of simulation results (Figure 6). The original layout was based upon a general knowledge of the process and the structure and was laid out with a wide range of geometries in an effort to compensate for lack of process level simulation and expected process variation. These were laid out as four terminal devices, with an independent body contact and placed in a pad frame as shown in Figure 7.

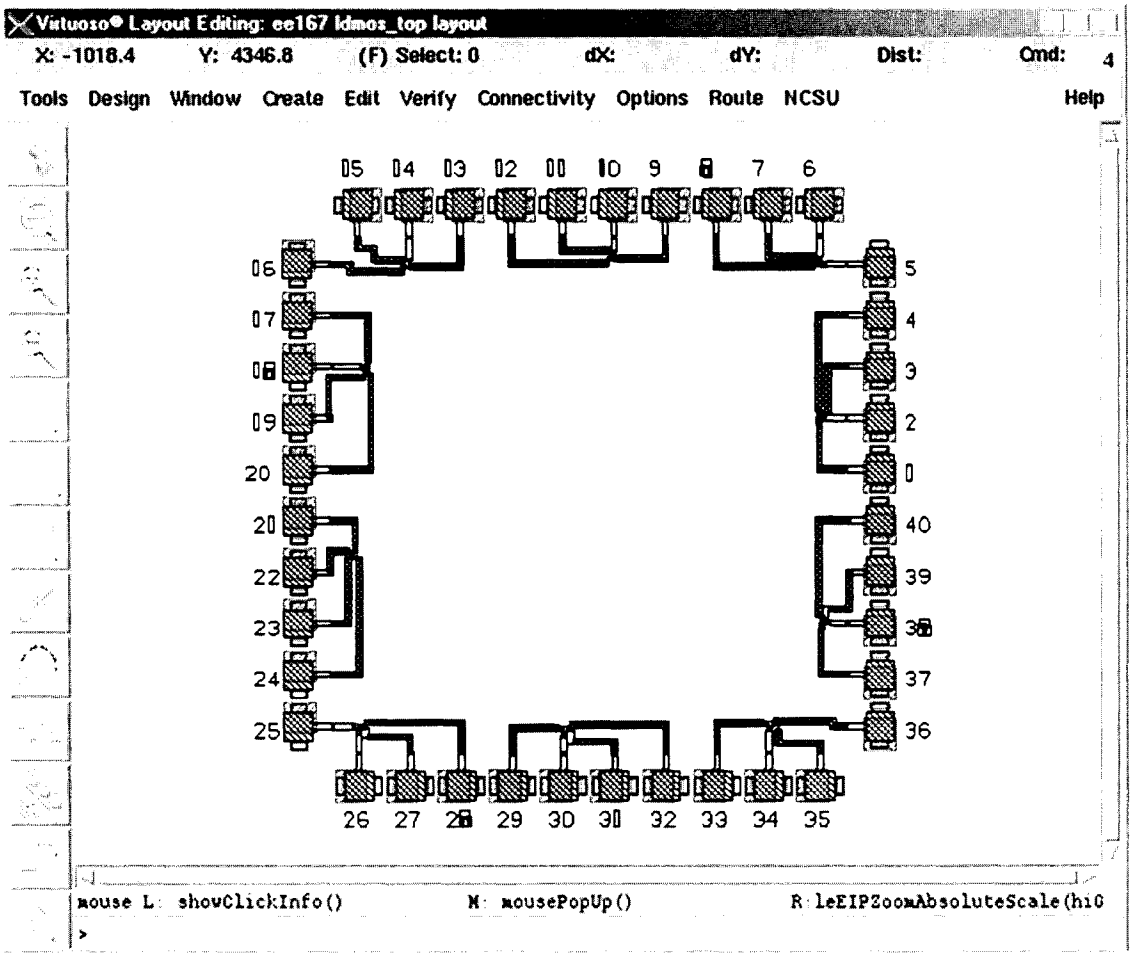


Figure 7. Padframe in Cadence Virtuoso Layout

This cadence layout file was used to generate a GDS2 file which was used to generate the Maskviews layout file, a product to manage layout and masking information compatible with Virtual Wafer Fab (VWF), by using GDS2 stream output (Figure 8).

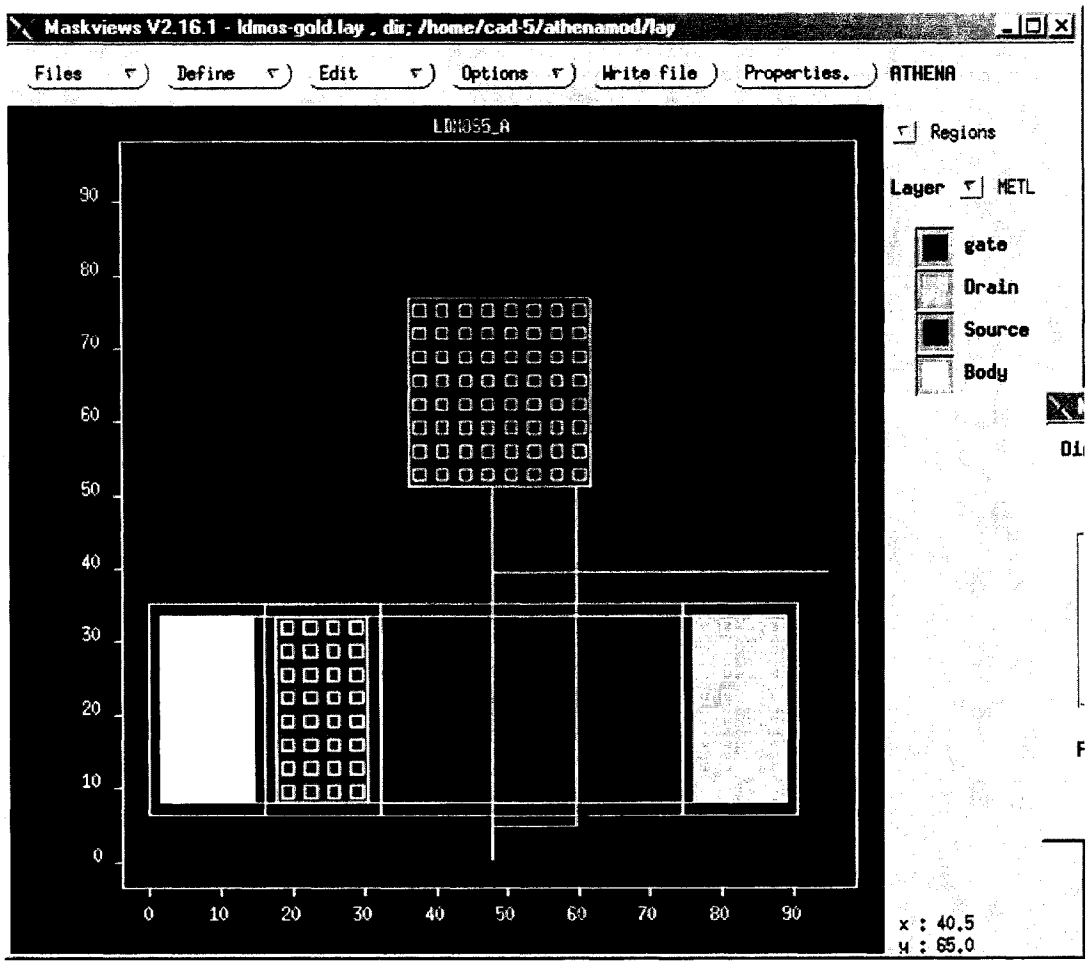


Figure 8. LDMOS Imported into Silvaco Maskviews.

Maskviews was an indispensable tool for this type of simulation. Once the layout file was imported, significant process simulation control become available. For example, it was possible to define the simulation grid and pick cut lines through various sections of the functional device. This greatly simplified the simulation of the device because "generic" run decks could be employed. The second version of the layout is shown in Figure 9, where the mask definitions are as follows: METL is the metal layer mask definition, CONT is the contact layer mask definition, NSEL is the n-select implant

window layer mask definition, PSEL is the implant window layer mask definition, POLY is the polysilicon layer mask definition, ACTV is the gate oxide layer mask definition, and NWEL is the well layer mask definition. These naming conventions are standard thought this document.

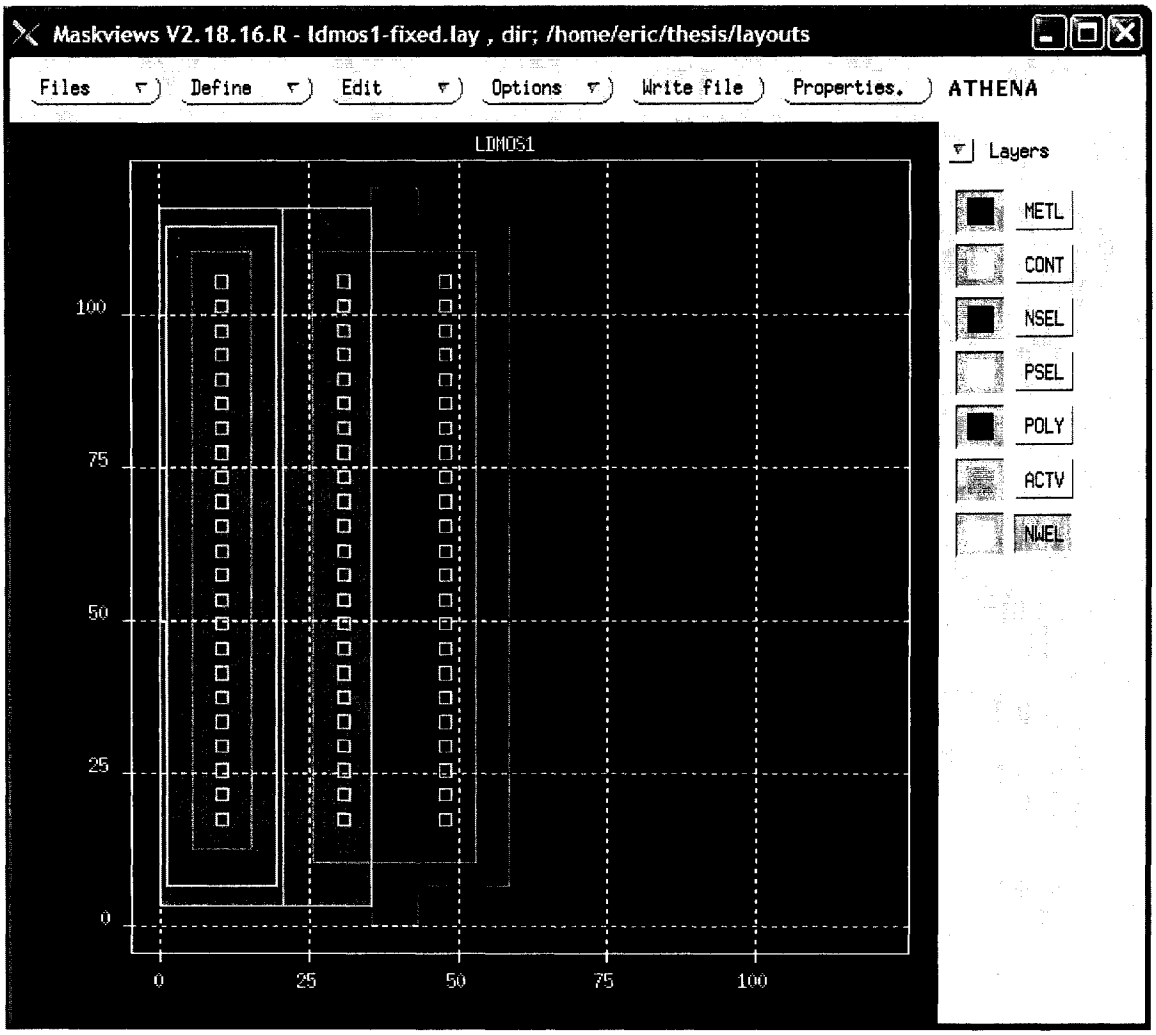


Figure 9. LDMOS Layout from Silvaco Maskviews - Rev. 2.

Generally the recommendations for a generic rundeck included removing position etch and definition statements and using regions defined by mask views instead. Defining the regions for processing was simply accomplished by calling the mask level of interest. This way, the same rundeck with varying cutlines was used to understand in detail a particular set of features rather than simulating the entire device for each run. The masks that were used for the 2d simulation are shown in Figure 10.

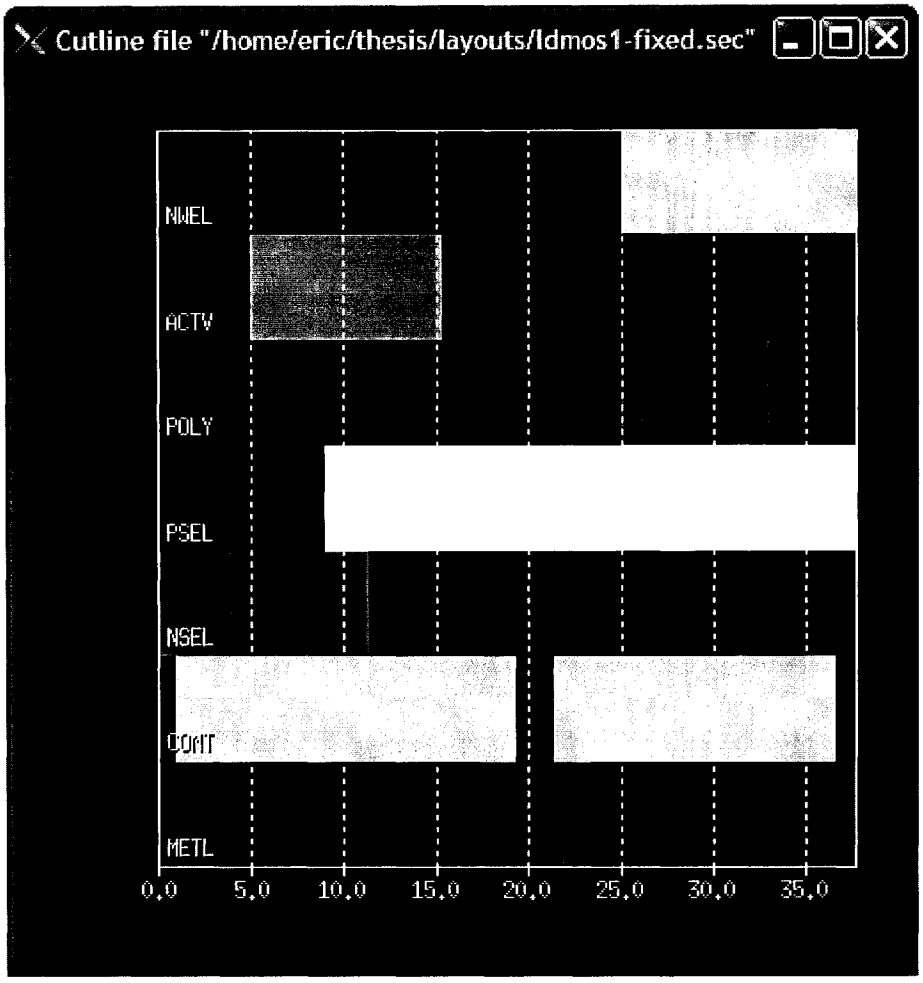


Figure 10. LDMOS Z-Axis Cutline. Figure Generated using Maskviews Showing the Correct Polarity for Clear Field and Dark Field Masks.

It is important to note that the Silvaco software is a challenge to learn and use appropriately. It is common for the manuals to contain contradictory information and features which are discussed display other than predicted results during the simulation. For that reason the Rundecks for this simulation were extensively documented as part of the code, and referenced to the traveler as well. The intention was to provide these rundecks as a learning tool for other devices, which may be constructed at the Spartan semiconductor fabrication facility. Toward this goal, devedit and adaptive meshing was used to handle the detail of correct meshing for the regions of the transistor. Ideally, loading a new cutline from Maskviews input downloaded from the GDS2 file form Cadence Layout would form the masking file. Then meshing (a very important part of correct modeling) would occur using the meshing utilities built into Athena. The rundecks were generic, and included features to clearly outline the process deviations from run to run. An example of the rundeck is shown below. The process simulation file (Figure 11) and the process traveler (Figure 12) are directly cross referenced.

```
============================================================
==
# ========== 4.1 MASK 2 Open an active area for etch

mask name="GOX"
structure outfile=4-1.st
tonyplot  4-1.st

# ============ 4.5 Etch Active Area
# Active Area definition etch

# Oxide Etch in BOE bath
rate.etch machine=BOEbath oxide a.m  wet.etch isotropic=371
etch machine=BOEbath time=12 minutes
structure outfile=4-5.st

# ============ 4.7 Strip the photresist
strip
struct outfile=4-7.st
```

Figure 11. Silvaco Rundeck Code Snippet.

PROCESS SESSION #4 Active Growth (Gate Oxide Growth) **DATE**

#	Time Est	Procedure	Parameters	Date	Cmpt Time	Sign	NOTES
4.1		Expose	MASK 2 - 10 second expose • Alignment is critical				This mask defines the Active area for the NMOS and PMOS transistors The FOX is etched away in the area defined by the mask
4.2		Develop	OCG 3:2 developer				
4.3		Inspect	Minimum feature size top and bottom CD=3microns, check alignment structures				
4.4		Hard Bake	• Set PL oven for 130°C • When at set point load wafers quickly • When the oven temp come back to 130°C, post bake for 30 minutes • Remove wafers • Transfer to a cool cassette				
4.5		Etch Oxide	• Use fresh Buffered Oxide Etch • Determine etch rate with one process wafer • Agitate • Dump Rinse				RIE is an alternative etch process
4.6		Inspect	FILMETRICS • Verify Oxide is etched away • Calibrate with blank				
4.7		Strip PR	• 15 min hot (T = 110°C) H2SO4 (75%) + H2O2 (25%) Piranha • Dump Rinse				
4.6		Spin	120Seconds Under N2				

Figure 12. Process Traveler Excerpt.

The extensive cross referencing of the traveler and the Rundeck was partially intended to supply a rundeck/traveler set as a teaching tool and also so that process variation may be modeled on the fly and outcomes predicted so that corrective action could be taken. The cross referencing along with the numerous plot structures generated as part of simulation allow students to observe the development of the semiconductor device as it is being processed. This example above also illustrates the use of fab specific processing equipment. In this case, the BOE bath etch was a characterized and used in the rundeck to show the slight undercut which occurred as a normal part of processing. This is a classic example of the subtle power of the commands in the rundeck as extreme over etch may not only distort the surface, but will also destroy the simulation mesh. In the case of our process, the variations in the etch process and the etch rate of the oxide resulted in etch until clearing was measured in the test structures. However, etch gradients are unavailable in Athena, so etch was simulated continuously, rather than interrupting the simulation. Since it is often mentioned that considerable expertise is necessary to run the simulation tools correctly, discussion of the use of commands was also included in the rundeck.

Another valuable use of generic rundecks is the availability of relatively easy misalignment experiments. As presented in the results section, several experiments were run as a function of misalignment data collected in lab. This was relatively straightforward using the "misalign" command.

Fabrication Process

Due to the detail contained in the process travelers, only a brief processing outline is provided here. N-type (P= $1*10^{14}$, <100>) wafers were oxidized in the furnace to create an implant protection oxide, and then implanted (Do=$1*10^{11}$, 140KeV, SP=P, Tilt=7) to standardize the surface doping concentration. Mask 1 defined the P-wells of the CMOS device by protecting the areas that remained N-well (since it was initially an n-wafer). Wafers were implanted (Do=$3*10^{13}$, 100KeV, Sp=B11, Tilt=7) and then etched in BOE to remove the protective oxide. Wafers were wet oxidized in the furnace to form the field oxide and then a drive-in step was performed defining the channel width in the asymmetric devices and well depth for the self aligned CMOS devices.

Mask 2 defined the active area for the transistors, and subsequently a 630 Å dry oxide was grown for the gate oxide. 5000Å of polysilicon was sputter deposited and then implanted (Do=$1*10^{15}$, 130Kev, SP=P, Tilt=7) to make the poly conductive. Mask 3 defined the poly lines. Mask 4 defined the P-Select source and drain regions of the PMOS which were opened over the common n-wells. The window included the poly gate, but since the boron implant (Do=$1.5*10^{13}$, 60KeV, Sp=Bf2, Tilt=7) did not penetrate the gate, only the areas to the left and right of the gate were doped. This formed the standard self-aligned PMOS transistors. Mask 5 defined the N-Select regions of the NMOS source and drain. Implant dose was (Do=$5*10^{15}$, 50KeV, Sp=As, Tilt=7).

Contact holes were etched by defining Mask 6 and etching through the FOX with Reactive Ion Etching (RIE). Metal was deposited by evaporating Al or sputter deposition

of an Al (1%Si) layer onto the wafers. Mask 7 defined the metal traces the wafers were low temperature annealed in forming gas at (450°C). Once removed from anneal, the fabrication processing was completed and the devices were ready for testing.

Test Procedure

Once the wafer fabrication was complete, the LDMOS devices were tested on the probe station in the Spartan Semiconductor Fabrication Facility. Device testing consisted of determining the DC transistor characteristics using HP4156A Semiconductor Parameter Analyzer, a custom probe card from SemiProbes and Signatone S-1150 probe station. The instrumentation was controlled by Metrics Technology Interactive Characterization Software through a GBIB interface. One important note, the original 40 pin probe card and pad system shown in Figure 7 was too frail to stand up to repeated testing. The second generation of devices used a four pin probe card and direct connection through alligator clips to the source monitor units (SMU) units of the 4156A. This standard pad frame and probe card is used for most of the device testing at San Jose State University.

Results

Simulation

The results of the LDMOS simulation were atypical for a LDMOS device. Typically in DMOS devices V_t is highly variable and difficult to control, while the V_{dsat} is relatively constant and high due to the short channel length as a result of the double diffusion (Figure 13). Because of the independent alignment of the poly gate to the mask prior and the NWELL being the first mask, and thus not aligned to any feature, there was considerable variability in the poly line to NWELL alignment. As a result of this and the fact that a long diffusion to create the PWELL is also functionally the channel, the dopant level in the channel is much lower than would be expected of a LDMOS and the channel width much larger. In the case of the Spartan Semiconductor process, the channel concentration did not vary much in simulation (see Table 3). As is clear from Table 3, the simulation predicted a shift of less than a volt across a misalignment of +/- 1.5 μm. This concurs with what would be predicted from theory because if the channel surface concentration does not vary greatly, neither will the V_t. This is also shown in the V_t graph (Figure 13). As one would expect, however, there was a significant difference in saturation current as a function of the channel length. This coupled with the change in V_t predict noticeable differences in the I_d vs. V_{ds} curve as a function of misalignment characteristics. The I_d vs. V_{ds} curve in (Figure 14) is illustrative of the results.

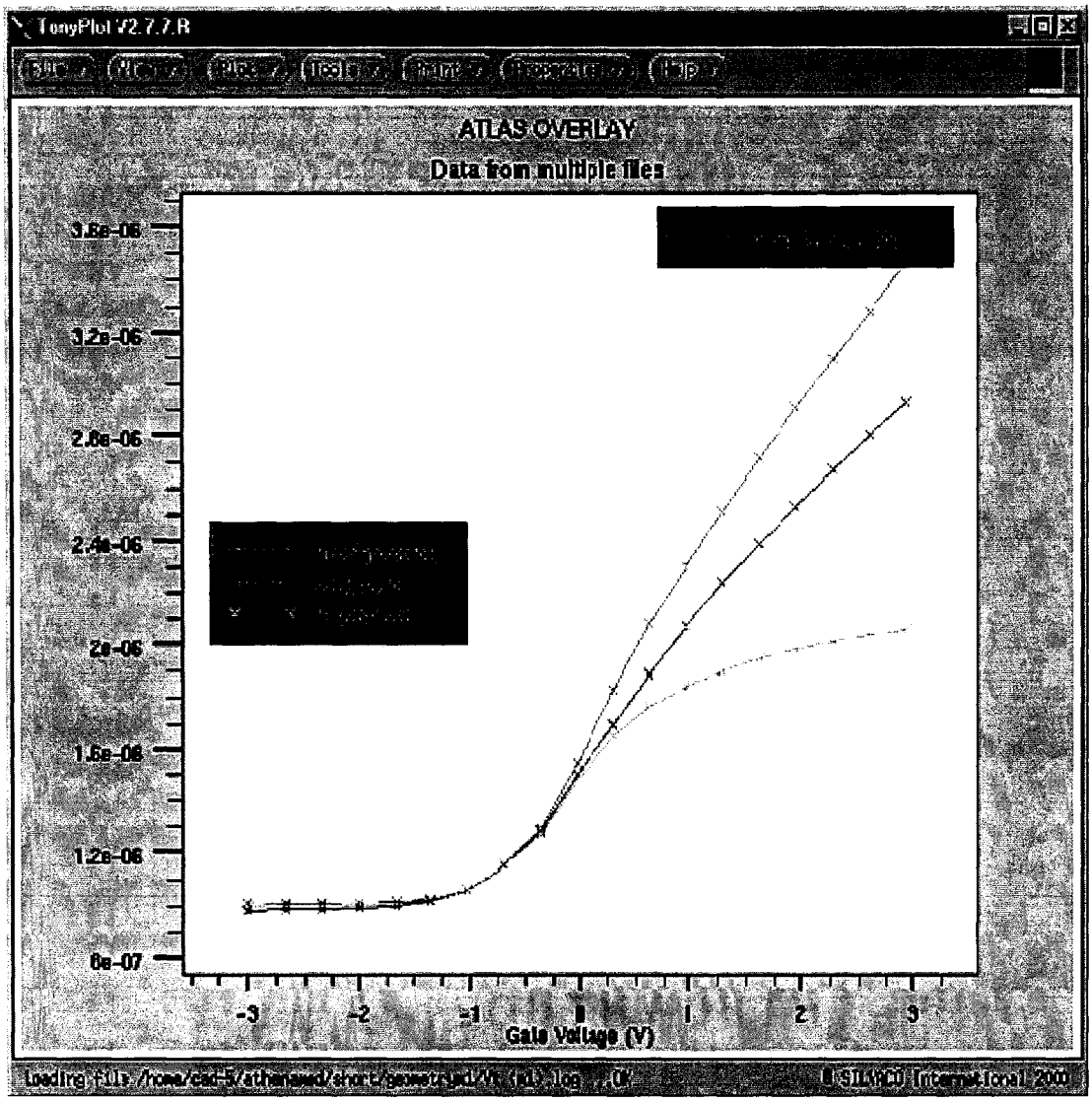


Figure 13. I_d vs. V_{gs} as a Function of Misalignment. Upper Curve (p1) and Lower Curve (m1) Lines Refer to the Shift of the Mask Alignment to the Drain or the Source.

Table 3. Simulation Results

	Correct Alignment	Drain Shift +1.5	Source Shift -1.5
Screening Oxide= angstroms			
X.val=0.4	1055.73	1055.73	1055.73
nxj=um from top of first Silicon layer			
X.val=0.4	0.352347	0.350257	0.293609
nxj=um from top of first Silicon layer			
X.val=0.4	2.72147	2.78024	2.6618
chan surf conc X.val=3 atoms/cm3	3.95E+14	4.67E+13	8.60E+14
nvt=	-2.07005	-2.34002	-1.49522
nbeta=	7.33E-06	6.38E-06	1.05E-05
ntheta=	0.0530589	0.122481	0.0802591
nidsmax_Vsb0=	6.86E-05	4.69E-05	8.88E-05
sat_slope_Vsb0=	4.74E-06	3.79E-06	5.79E-06
nidsmax_Vsb-2=	7.50E-05	5.73E-05	0.000100202
sat_slope_Vsb-2=	2.39E-06	1.57E-06	3.64E-06
nidsmax_Vsb2=	4.65E-05	2.84E-05	6.59E-05
sat_slope_Vsb2=	9.08E-06	8.95E-06	9.42E-06
Field oxide=angstroms X.val=0.4	3546.11	3599.38	3599.38
gateox= angstroms X.val=3	630.198	630.198	540.099

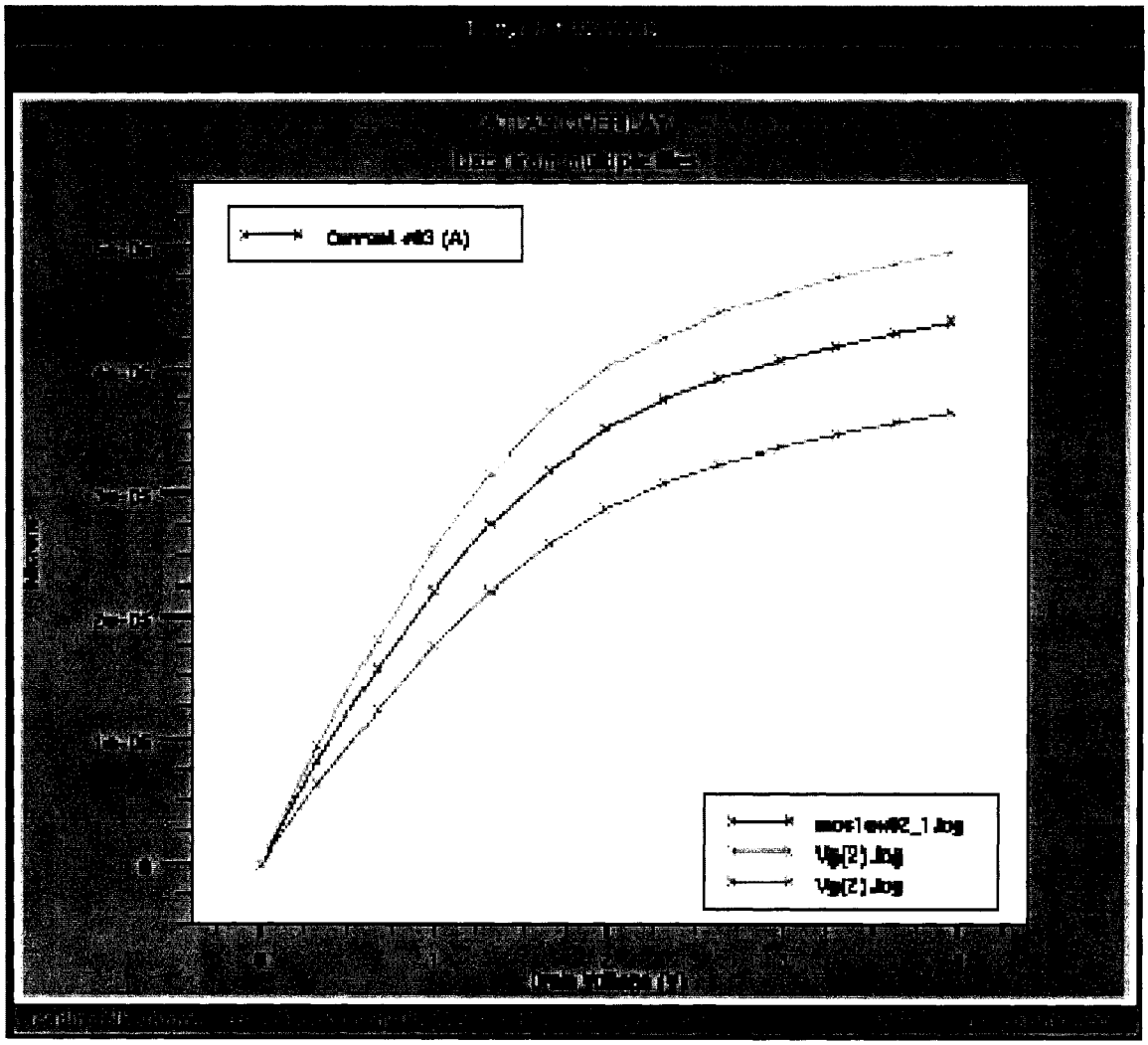


Figure 14. I_d vs. V_{ds} as a Function of Misalignment. Upper Curve Results from a Drain Shift, the Lower Curve Results from a Source Shift. The Middle Curve is the Nominal Aligned Structure.

As is stated in the Silvaco instructional manual, the majority of the time spent was in creating an accurate Athena file, and then many Atlas files were run in parallel. One of the most challenging aspects of simulation is correctly meshing the device. Too loose a mesh fails to capture subtle effects or grossly misestimated them, such as a dense grid resulting in an overly high current through the channel. However, the processing time required increases geometrically with the increase in mesh triangles.

Fabrication Results

The fabrication results for the first generation devices were certainly lower than expected. Several factors contributed to this, but the primary factor was the lack of statistical control of the operating procedures in the fabrication facility. An excellent example of this was an unavailability of cooling water for the spring term. This resulted in extreme temperature swings, which in turn resulted in erratic equipment performance and increased maintenance requirements. Another unexpected result was the inconsistency of the thermal expansion coefficient of the borosilicate masks and the substrate. The result of this was the necessary "averaging" of the misalignment across the wafer, i.e. the wafers would be misaligned to the left on the left side and misaligned to the right on the right side. This effect, taking into account user error, random variation and the result of simulation illustrates that the process should be altered to align the poly gate to the NWELL rather than the Gate Oxide mask for optimum LDMOS repeatability and reliability. This simple alteration should not have any effect on the fabrication of

standard CMOS devices and is the direct result of understanding the process variation and comparing it to variation in the simulation as a function of misalignment.

The lack of a high enough contrast ratio in the photolithography process also presented challenges. As a result the wafers were reprocessed several times during the contact photolithography step because adequate contact holes could not be defined in the photoresist. In the case of some wafers, this resulted in completely device failure. Prior to mask layout, gamma curves to determine control charts for minimum and maximum resolution would have mitigated this issue, and have since been incorporated into standard lab operating procedures.

Another challenging effect was the lack of access to a reliable, portable padframe. Pad frames contain the input and output pins and typically a buffer and ESD protection devices. Issues with the ESD devices are common in uncharacterized processes because they are essentially only diode structure with voltages exceeding the operating design point of the circuit. In this case, testing showed that the diodes turned on at slightly greater than a volt (4 volts less than the operating point of the design specifications) and had pad-to-pad substrate conductance on the order of 330Kohms. This resulted in some rather confusing test results. Fortunately a remedy was available and a mask was generated to remove the metal contact to the ESD protection devices. The second set of masks did not incorporate any ESD protection devices. This is reasonable in a research facility as long as users are careful to ground themselves before testing commences.

The issue with the diodes in the ESD protection circuit actually illustrates a rather critical point in device fabrication. Early in the semester test structures were designed

according to the available literature, but failed to include simple diode structures and MOS structures. These devices would have assisted in the analysis of the quality of the gate oxide and the presence of substrate leakage currents. Both of these turned out to be issues in the process. Concerns about the oxide quality were due to the exhaustion of the nitrogen source during the gate oxide anneal – which resulted in aberrant oxide thickness. This indicates that oxidation may have occurred in the anneal stage resulting in and increased trapped oxide charge and lowered V_t.

As a result of using industry standard design and layout tools, the next series of LDMOS benefited from design and evaluation of the test structures. In addition, the process traveler was meticulously documented with explanation of the in process actions in the margins and cross referenced to the rundeck so that less process variation due to training issues should occur in the future. Indeed, the second set of devices showed much greater success rate and produced an acceptable yield. An example device from the second generation device is shown in Figure 15.

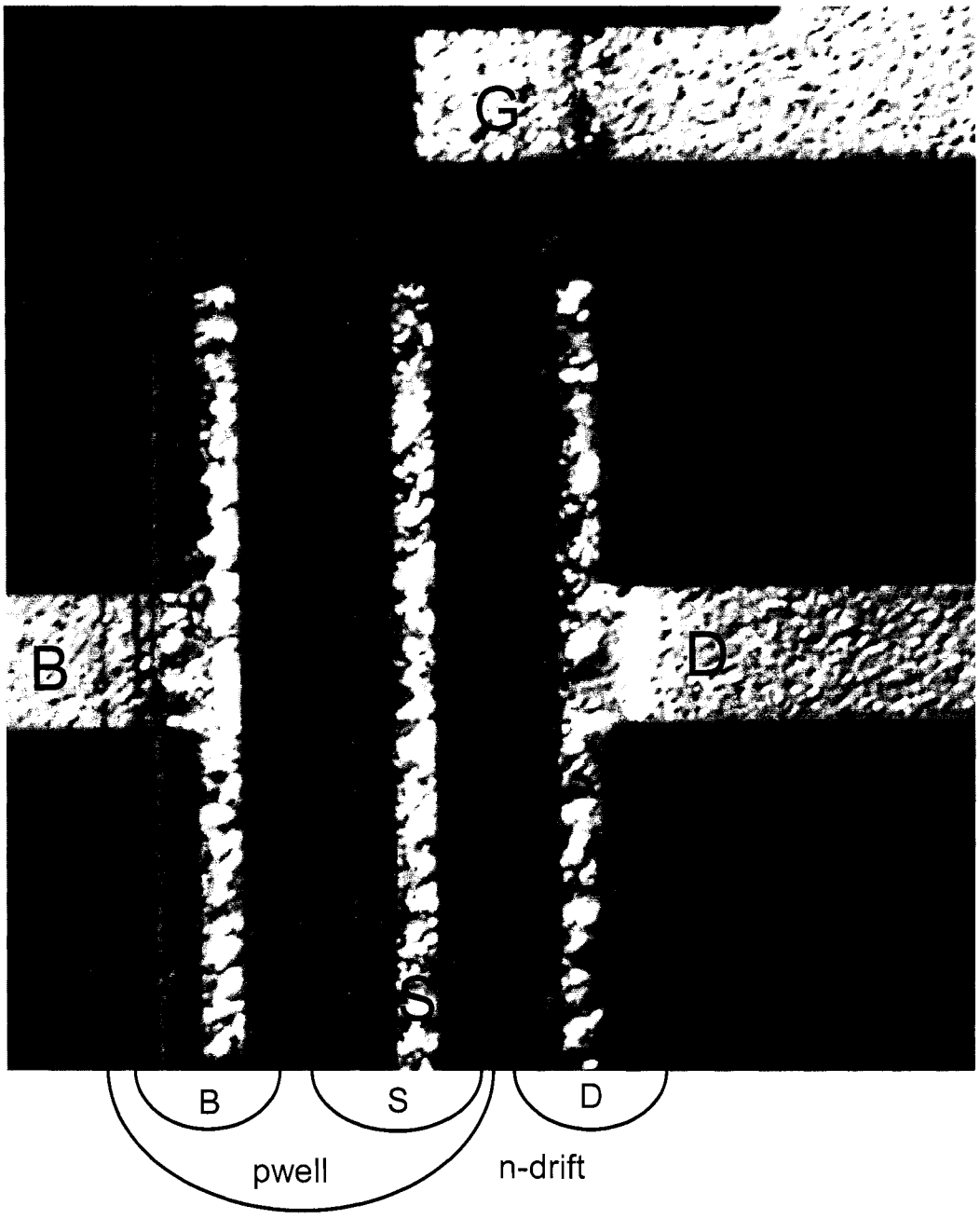


Figure 15. Optical Photomicrograph of Fabricated LDMOS Device. The Nominal Z Axis Device Structure is Shown With the Line Drawing below the Photomicrograph for Reference Purposes. Terminals are Labeled. Pads not Shown.

Some notes on the fabrication of the asymmetric (LDMOS) devices are in order. To form the asymmetric device the Pwell was aligned to the edge of the poly gate. Thus, as stated in prior sections, these devices were not self aligned. The Pselect was used to form the body contact in the pwell, next to the source. The source and body contacts were junction isolated. Figure 15 shows the implant windows with a diagram which aligns to the features in the micrograph to clarify the structure of the LDMOS. Since the alignment sensitivity of our process is on the order of 1μm, devices must be carefully selected which have the optimal alignment. In this case, two wafers were selected which had optimal alignment between the critical layers. These wafers were used for characterization.

Device Test Results

There were significant processing issues with the first generation of devices. Many of the wafers appeared to have a misty layer of aluminum. Only a couple of the wafers had areas free of this defect that could not be removed by etching. Of those two wafers, one was tested to determine the performance of the DMOS devices (wafer B2). A standard threshold voltage test (using a 4 probe parameter analyzer) was run on each of the devices in the two dies sampled (a total of 20 devices) (Figure 16).

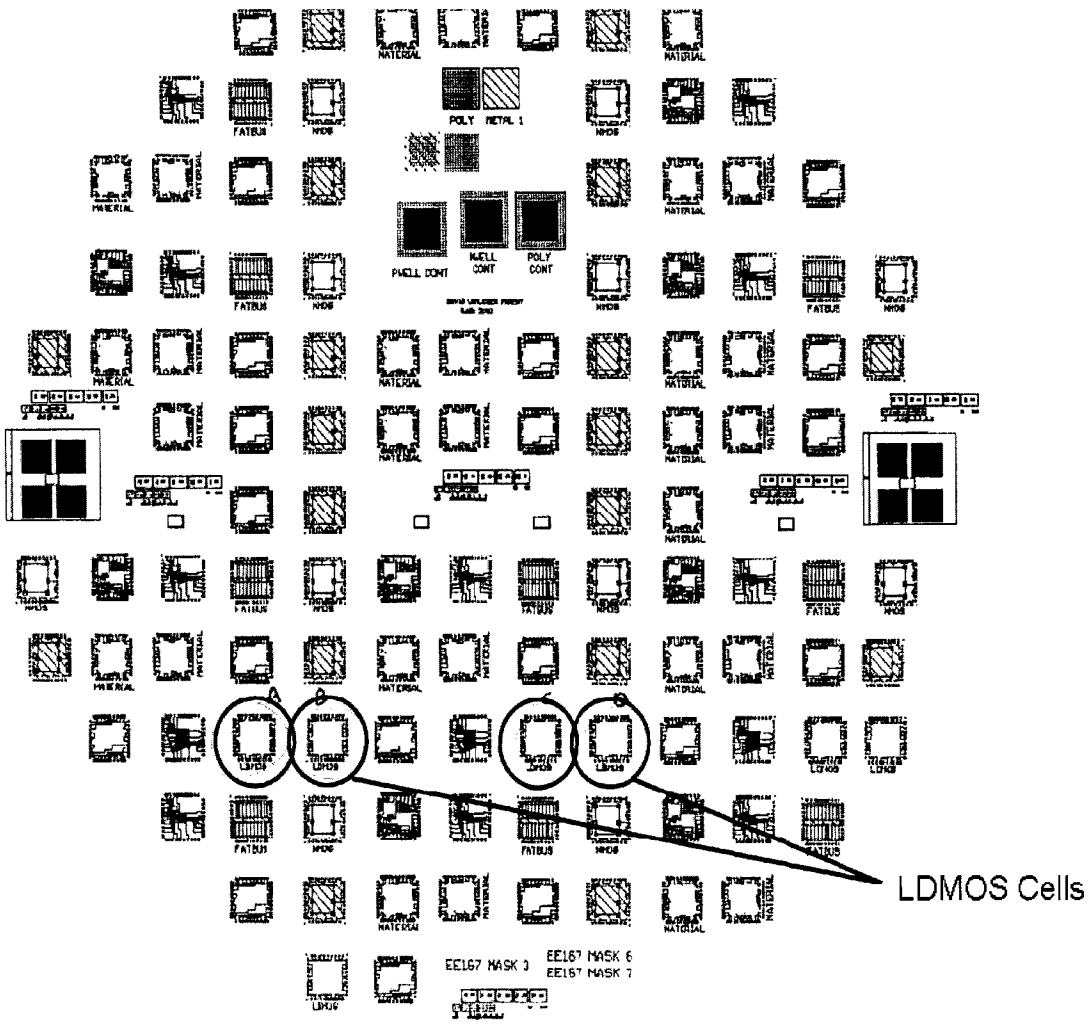


Figure 16. Wafer Map of Rev. 1 Devices.

A typical device appeared to have a V_t of about 0.325 V (Figure 17).

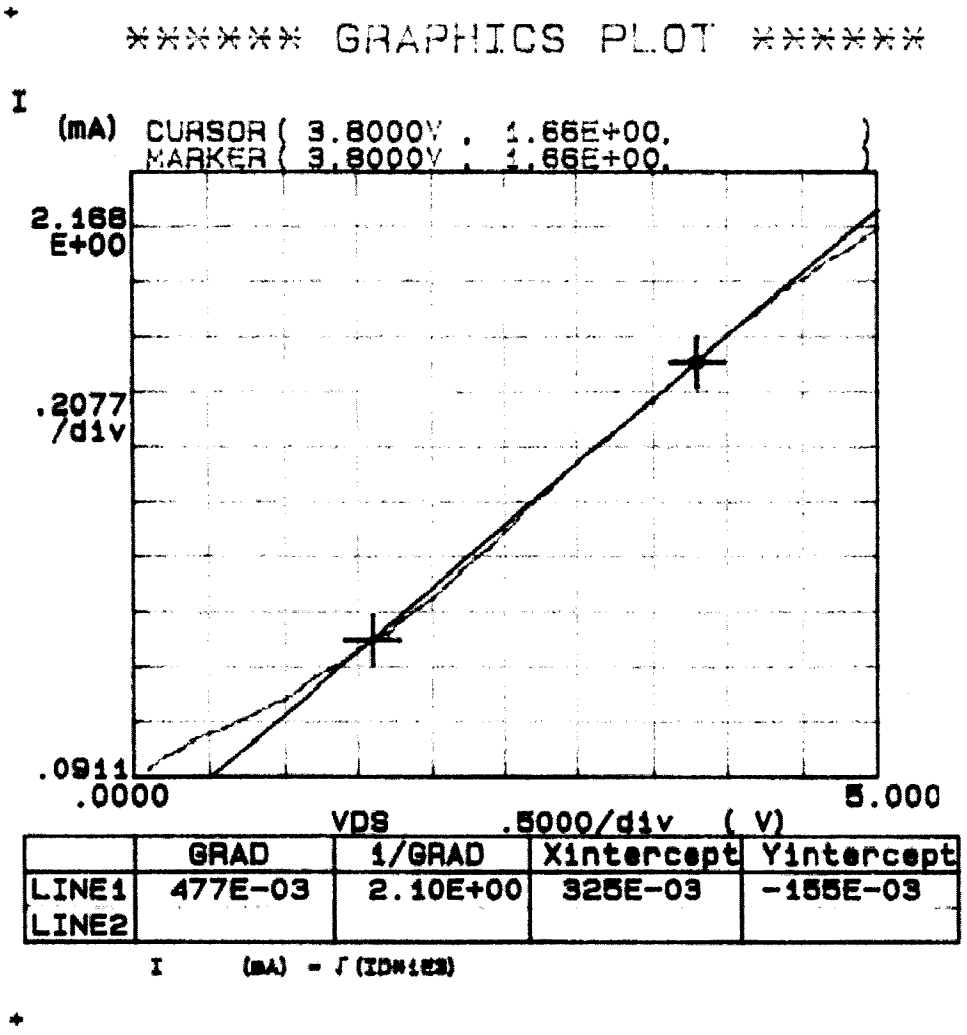


Figure 17. Single I_d vs. V_{gs} Curve Showing V_t.

An I_d vs. V_{ds} curve was performed which revealed a strong consistency among the devices, but which was not FET-like in appearance (Figure 18).

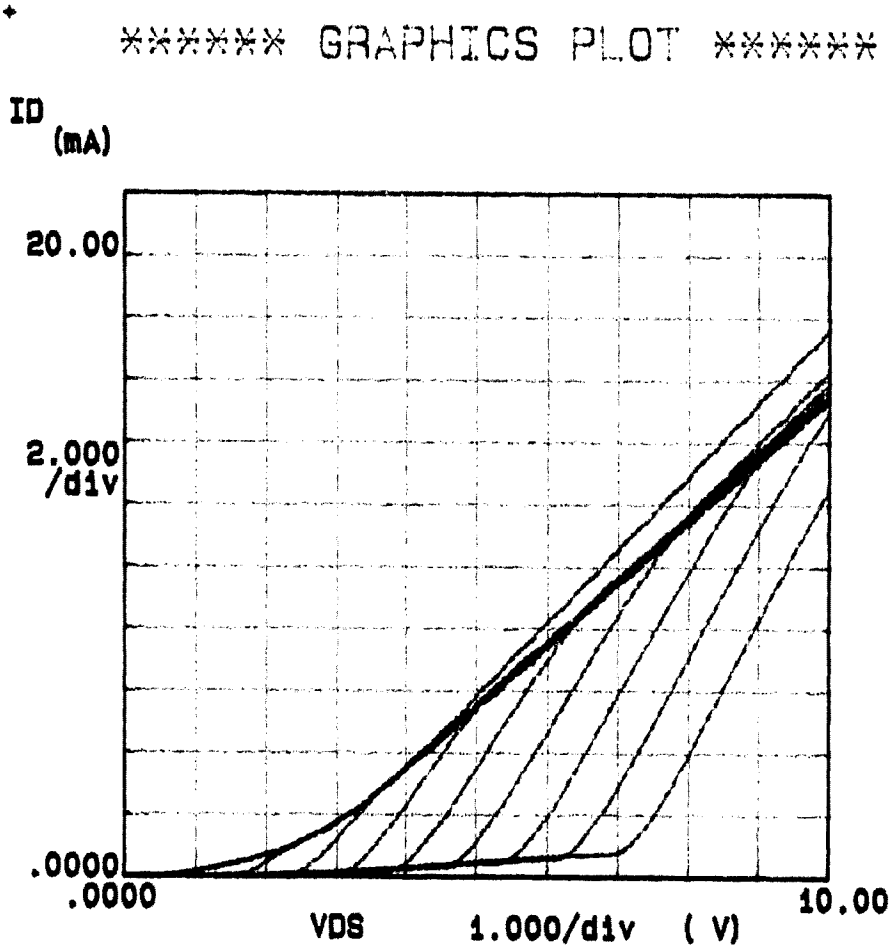


Figure 18. I_d vs. V_{ds} Curve with Stepped V_g.

It appears that the most likely cause of difficulty is the padframe that the devices were mounted in. It was suggested that there may have been a short in the ESD protection diodes on the padframe, which turned out to be correct. Additional testing of this idea

demonstrated diode-like I-V response between unconnected pads, which would appear to verify this initial assessment.

The second generation of devices was much more successful. An error in the pwell doping implant recipe resulted in a LDMOS channel doping that was an order of magnitude to high. This had two significant effects. First, the threshold voltage of the devices was near 6V. Fortunately, the oxide was high enough quality that gate breakdown did not occur until near 25V – 30V for the small number of devices that were destructively tested. Second, the high channel doping resulted in a longer channel width than expected and less of a retrograde well effect than was predicted form simulation.

Even with these caveats, the devices were relatively straightforward to test and interpret demonstrating the robustness of the extraction methodology and the modeling approach. The most direct way to demonstrate the asymmetric nature of the LDMOS devices is to observe the results of the I_d vs. V_{ds} curve (forward operation) and I_s vs. V_{sd} curve (reverse operation) curves (Figure 19).

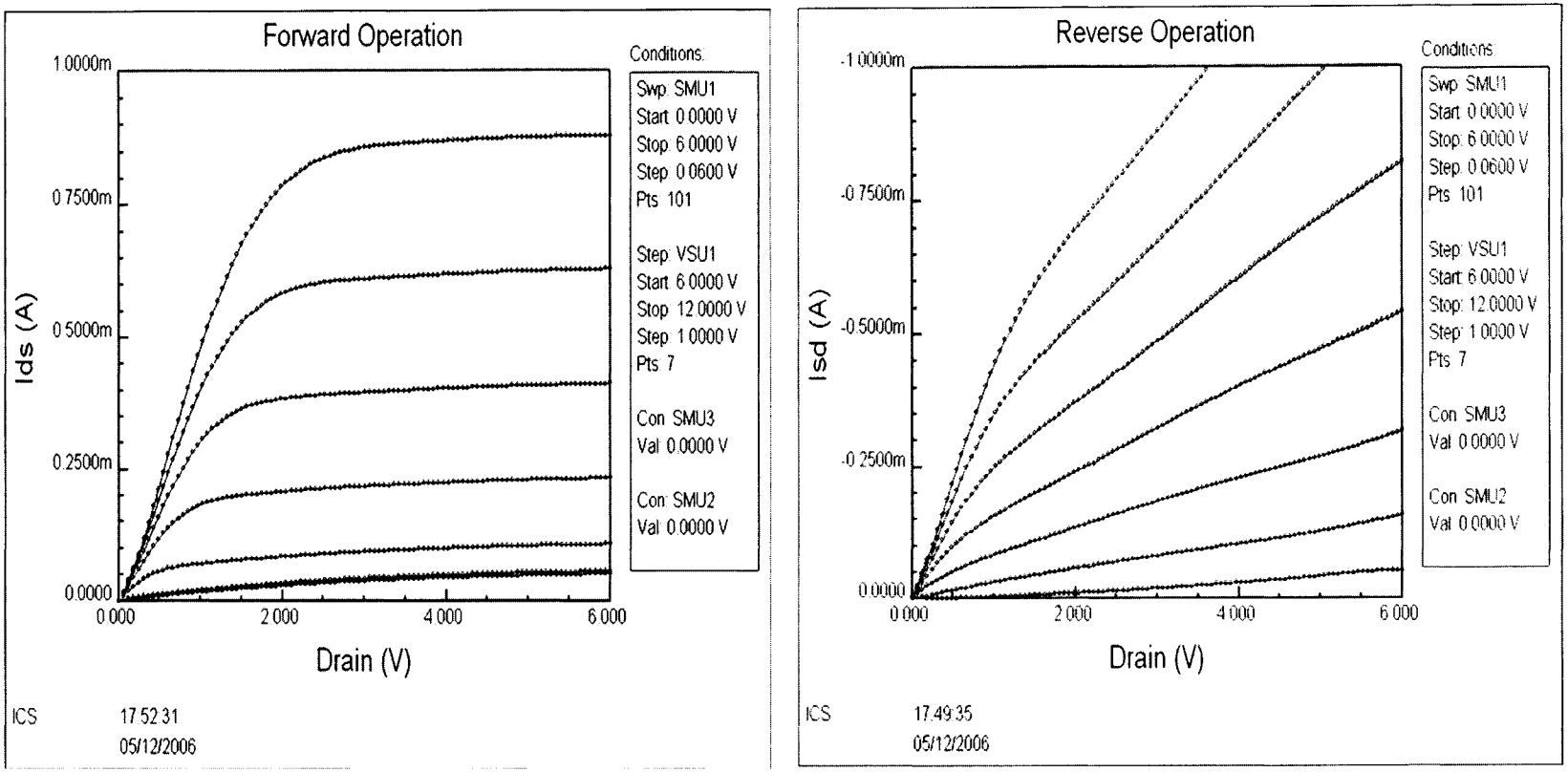


Figure 19. Forward and Reverse I_d vs. V_{ds} Curve, Stepped V_g.

Comparison of the curves shows that for reverse operation the initial channel current at the onset of saturation is significantly reduced and the channel length modulation effects increase significantly. This is consistent with the theoretical predictions because the doping is much higher on the source side and there is no drift region next to the channel. Therefore, the depletion region extends deeply into the channel region in reverse operation as a function of increasing source voltage. Since the channel is asymmetrically doped and can be approximated by a set of series transistors with differing turn on voltages as function of the channel doping [87], the observed change in the initial turn on current can be posited as well. The source side transistor will always have a higher turn on voltage due to the doping concentration being the highest in the channel on the source side. Under forward operation the "end" transistor on the source side will be the last to turn on fully, while the virtual transistors in the channel are already fully conductive. In reverse operation, the drain "end" transistor would turn on first. This would appear as a virtual cascode, which will have a longer channel length, and thus lower current drive capability.

The next significant metric is the extraction of n. This requires the determination of I_o, the technology current, which is required for analysis nonetheless (Figure 20).

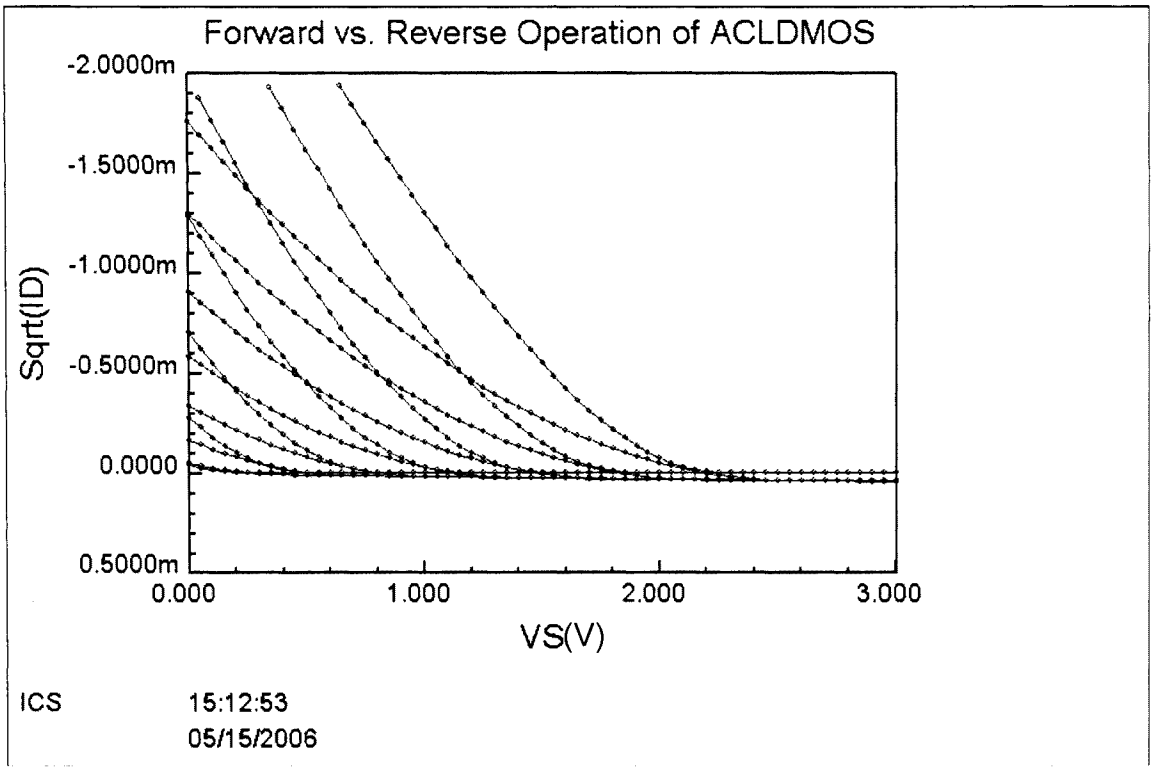


Figure 20. I_o Extraction Curve. The Set of Six of Curves with a Steep Slopes is Forward Operation, The Set of Six Curves with a More Gradual Slope is Reverse Operation

Once again notice the strong difference between forward operation and reverse operation. The technology current Io is extracted as [52]:

$$Io = \left[2U_T(slope)\right]^2$$

Thus a steeper slope indicates a larger technology current. As shown in Figure 21, sweeping the gate voltage and measuring the source voltage (V_s) at a fixed drain current equal to the I_o results in a very useful analysis figure. From this single measurement, the

threshold voltage can be determined from the intersection of the curve with the V_s axis, as 5.4074V as shown in Figure 21.

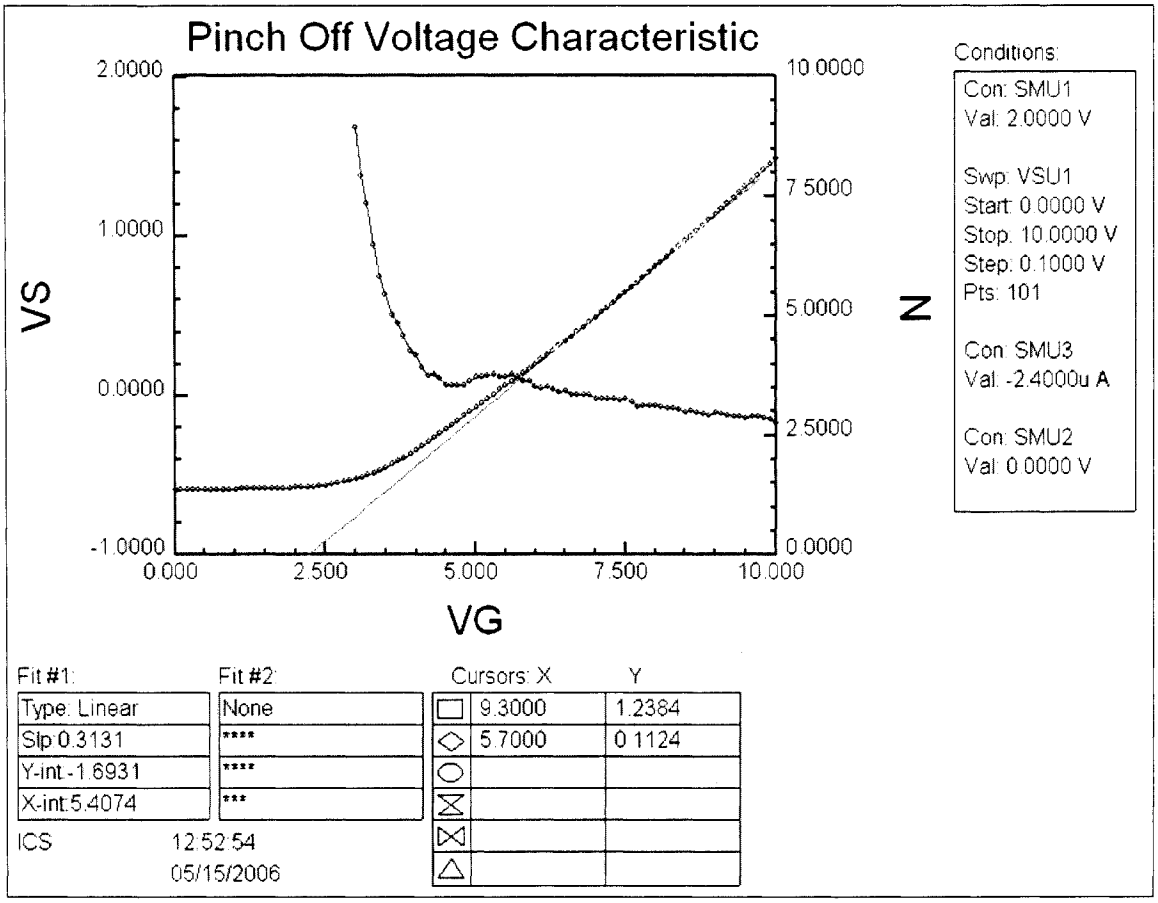


Figure 21. Extraction of n, V_{to} and φ.

The parameter n can be computed as:

$$n \equiv \left[\frac{\partial V_p}{\partial V_G}\right]^{-1} = 1 + \frac{\gamma}{2\sqrt{V_p + \varphi}}$$

Where the measured V_s is equal the pinch off voltage, V_p and V_g is the gate voltage, γ is the substrate effect factor; φ is the approximation of the surface potential in strong inversion. At the lowest end of the subthreshold curve, n rises rapidly. This is most likely due to the high electric field at the surface. The electric field affects the carrier mobility according to [4]:

$$\mu_{surface} \propto \varepsilon^{-\frac{1}{3}}$$

This can be confirmed with by measuring the mobility reduction effect (Θ) on a I_d vs. V_{gs} graph with the same technique used to determine this parameter in a SPICE level 3 model – albeit with the precisely determined V_{to} from the V_p vs. V_g graph. This metric has the added benefit of providing KP= μC_{ox} to provide info on the mobility. The clearer way to illustrate these effects is by referring to the transconductance efficiency figure of merit curves (Figure 22).

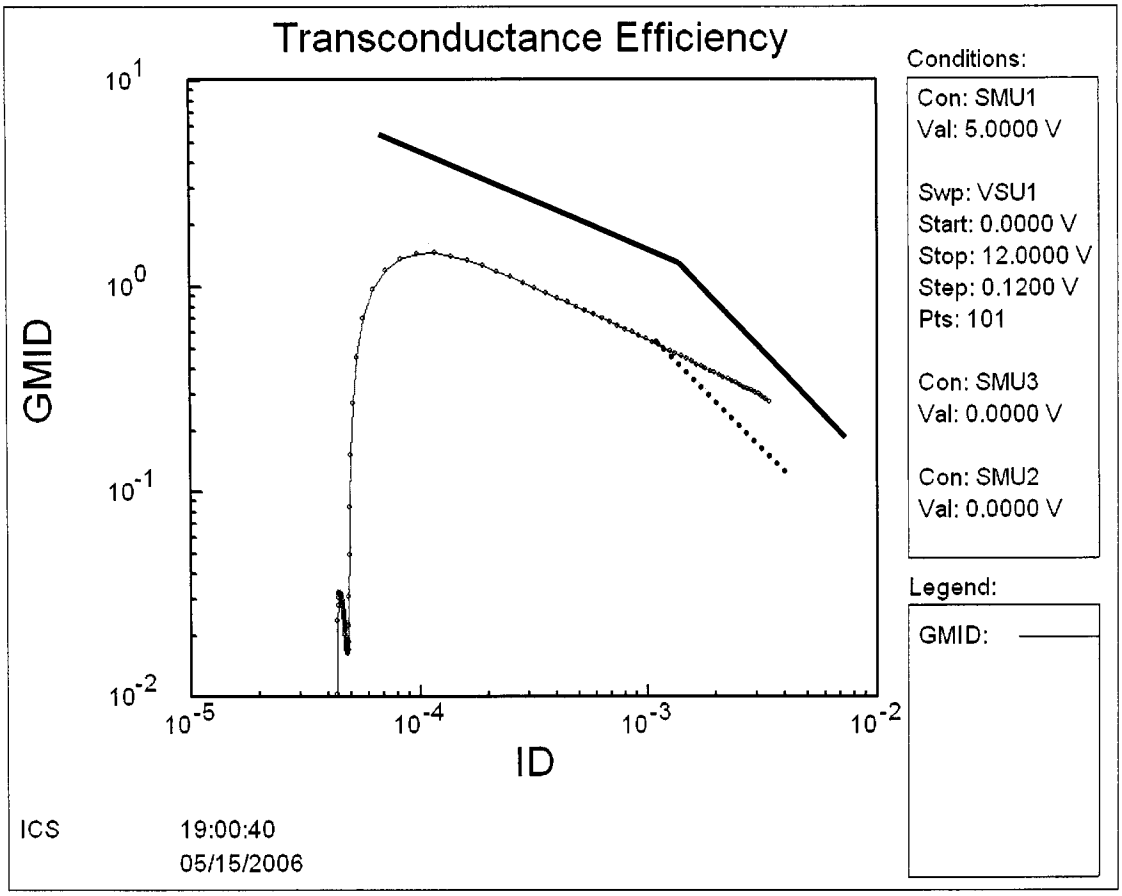


Figure 22. Measured vs. Theoretical Transconductance Efficiency Curves.

As is seen in Figure 22 the transconductance efficiency of this device can clearly be measured without detailed information about the processing details. The bold black line shows the theoretically predicted slopes of the strong inversion region (slope=-1/2) and velocity saturation (slope= -1) regions of operation. The dotted line curve was collected under conditions which would result in velocity saturation. The weak inversion operation

of this device was compromised by high substrate doping. The high doping would result in increased mobility field reduction effects and an increase in bulk capacitance, reducing n. This effect is easily observed on the left hand side of the graph and is mirrored in Figure 21. With simple information on the width and the length of a device, the quality of the process for analog circuit designs can immediately be evaluated. Even when the fabrication results are less than optimal, which is common in research processing technology development, the transconductance efficiency curve provides valuable insight and assists debug of the process. In the case of this device, atypical because it has an unusually high turn on voltage and low surface mobility, there is a clear definition of the movement from weak inversion, to moderate inversion, to strong inversion and finally reaching velocity saturation during operation. As predicted, the leakage current for forward versus reverse operation are not equal. As suggested in section III, the forward operation of an MOS device with and asymmetrical channel doping actually has a higher diffusion current, and thus the technology current is higher. Technology current is extracted according to [52] and shown in Figure 20.

Table 4 summarizes the design goals evaluated against a standard CMOS process for 2μm devices as used in the design example in the illustrated g_m/I_d presented earlier in this thesis. Significant gains were made with respect to channel length modulation effects, but the expected increase in technology current was not observed. Measurement of the transconductance efficiency figure of merit was hampered by a significant increase in the technology factor n. Note the forward operation technology current (I_o) is much higher in the forward direction than the reverse (Table 4).

Table 4. Test Results Summary of Fabricated Devices vs. Benchmark. Final Results for LDMOS Device Tested According to the Method Outlined. Arrows Indicate Desired Parameter Direction.

Metric (↓↑ goal)	benchmark	LDMOS
Is[A/u,l] ↑	50 – 200	73-75
Vea [v/um,l] ↑	4-8	100 - 150
n [.] ↓	1.3 – 1.5	3.5 @Is

BJT Operation

While these devices should have functioned admirably as a gate controlled BJT, the minimum base width was too long, and thus gains of 2 over a very narrow current range were recorded. In addition, the low gain is likely due to dopants, such as gold, commonly included in substrates used for MOS transistor fabrication included to reduce the carrier lifetime in the base.

Conclusion

Summary

In the introduction, scaling was shown to have severe impacts on analog transistor operation. The drive to optimize digital transistor operation has dramatic effects on analog transistor operation. A powerful analog design technique was outlined, and shown to be able to allow process designers to design transistors for a particular operation in circuit space. Using this technique, alternative devices used for power switching, LDMOS transistors, were hypothesized to have unique low power operational characteristics. This was verified by simulating and fabricating a CMOS compatible LDMOS device and subsequently testing the device using the level of inversion methodology. While results were not optimal, they showed the theory was correct and the testing method was very robust and did indeed lend insight into the minutiae of the operation of the transistor in low power regime. The brevity of Table 4 highlights the elegance of this evaluation methodology. Transistors can easily be compared for performance across process technologies, across transistor design styles and even across technology types (comparing non silicon substrates, for example). The reasonable number of parameters and the clear interaction between them allow direct insight into the operation of the transistor from a circuit design perspective.

Future Work

The next logical step is to use this initial investigation as a proof of concept platform for partnering with a semiconductor fabrication house. Ideally, one would expect that using these devices in deep submicron process would yield benefits in signal to noise ratio, reducing of the area utilized and improved gain per unit area. In this thesis fabrication was necessary to have control over the dopants and the exact processing so that simulation could be correlated to testing results. Often initial simulation is carried out without a benchmark process. While the results can be interesting, and useful for testing theory, communicating to a semiconductor fabrication house that they can see marginal improvement in performance for a significant process change supported by simulation data is problematic because of the risk involved. Process development and modern deep submicron fabrication is a complex, sensitive operation. The intent of this thesis was to investigate extended operational regions, and thus provide an added value to a device currently implemented in a commercially available process.

Bibliography

[1] P. E. Allen and D. R. Holberg, *CMOS analog circuit design*, 2nd ed. New York: Oxford University Press, 2002.

[2] J. J. P. Bruines, "Process outlook for analog and RF applications," *Microelectronic Engineering,* vol. 54, pp. 35-48, Dec 2000.

[3] E. Vittoz, "Present and future industrial applications of bio-inspired VLSI systems," *Analog Integrated Circuits and Signal Processing,* vol. 30, pp. 173-184, Feb 2002.

[4] B. Van Zeghbroeck, "Principles of semiconductor devices," Online: http://ece-www.colorado.edu/~bart/book/welcome.htm, 2004.

[5] A. Hastings, *Art of analog layout*: Prentice Hall, 2000.

[6] D. Foty and G. Gildenblat, "CMOS scaling theory - why our "theory of everything" still works, and what that means for the future," 2004, pp. 27-38.

[7] E. Vittoz, "Micropower Techniques," in *Design of analog-digital VLSI circuits for telecommunications and signal processing*, J. Franca and Y. Tsividis, Eds. Englewood Cliffs, N.J.: Prentice Hall, 1994, pp. 53-96.

[8] A. J. Annema, "Analog circuit performance and process scaling," *Ieee Transactions on Circuits and Systems Ii-Analog and Digital Signal Processing,* vol. 46, pp. 711-725, Jun 1999.

[9] A. J. Annema, B. Nauta, R. van Langevelde, and H. Tuinhout, "Analog circuits in ultra-deep-submicron CMOS," *Ieee Journal of Solid-State Circuits,* vol. 40, pp. 132-143, Jan 2005.

[10] C. Sodini and B. Wooley, "Technology Challenges for Mixed-Signal IC Design," 2003.

[11] B. Razavi, "CMOS technology characterization for analog and RF design," *Ieee Journal of Solid-State Circuits,* vol. 34, pp. 268-276, Mar 1999.

[12] M. J. M. Pelgrom and M. Vertregt, "CMOS technology for mixed signal ICs," *Solid-State Electronics,* vol. 41, pp. 967-974, 1997.

[13] R. R. S. Mudanai, W-K Shih, P. Packan and S-W Lee, "Halo Doping: Physical Effects and Compact Modeling," in *NSTI Nanotechnology Conference and Trade Show - Nanotech 2006 - 9th Annual*, Boston, MA, 2006.

[14] L. A. Akers, M. Holly, and J. M. Ford, "Transconductance degradation in VLSI devices," *Solid-State Electronics,* vol. 28, pp. 605-609, 1985.

[15] B. Murmann and B. Boser, "Digitally Assisted Analog Integrated Circuits," *Queue,* vol. 2, pp. 64-71, 2004.

[16] B. Boser, "Analog Circuit Design with Submicron Transistors," in *IEEE Santa Clara Valley (SCV) Solid State Circuits Society Monthly Meeting Presentation*, 2005.
[17] C. Fiegna, "The effects of scaling on the performance of small-signal MOS amplifiers," *Solid-State Electronics,* vol. 46, pp. 675-683, May 2002.
[18] B. Kleveland, C. H. Diaz, D. Vook, L. Madden, T. H. Lee, and S. S. Wong, "Exploiting CMOS reverse interconnect scaling in multigigahertz amplifier and oscillator design," *Ieee Journal of Solid-State Circuits,* vol. 36, pp. 1480-1488, Oct 2001.
[19] J. L. G. Xavier Aragonès, Antonio Rubio, "Substrate Coupling Trends in Future CMOS Technologies," in *PATMOS'97* Louvain-La-Neuve, Belgium, 1997.
[20] S. Donnay, G. Gielen, and W. Sansen, "High-level power minimization of analog sensor interface architectures," *Integrated Computer-Aided Engineering,* vol. 5, pp. 303-314, 1998.
[21] M. T. Bohr, "Interconnect scaling - The real limiter to high performance ULSI," *Solid State Technology,* vol. 39, pp. 105-&, Sep 1996.
[22] P. R. Gray, *Analysis and design of analog integrated circuits*, 4th ed. New York: Wiley, 2001.
[23] M. D. Hussein Ballan, *High Voltage Devices and Circuits in Standard CMOS Technologies*, 1st ed.: Springer, 1998.
[24] C. H. Diaz, "Bulk CMOS Technology for SOC," in *International Workshop on Junction Technology*, 2002.
[25] Y. Taur, "CMOS design near the limit of scaling," *Ibm Journal of Research and Development,* vol. 46, pp. 213-222, Mar-May 2002.
[26] S. Thompson, P. Packan, and M. Bohr, "MOS Scaling: Transistor Challenges for the 21st Century," *Intel Technology Journal,* vol. Q3, 1998.
[27] V. Zhirnov and J. A. Hutchby, "New Device Structures – Or, What Do We Do After CMOS?," *Future Fab Intl.,* vol. Volume 19, 6/28/2005 2004.
[28] T. Skotnicki, J. A. Hutchby, T. J. King, H. S. P. Wong, and F. Boeuf, "The end of CMOS scaling," *Ieee Circuits & Devices,* vol. 21, pp. 16-26, Jan-Feb 2005.
[29] J. Hutchby, V. Zhirnov, R. Cavin, and G. Bourianoff, "Functional Scaling Beyond CMOS," in *Interantion Conference on Solid State adn Integrated Circuit Technology (ICSICT)* Beijing, China, 2004.
[30] D. Flandre, "The gm/ID synthesis methodology: the missing link between symbolic analysis and design automation for MOS integrated analog circuits." vol. PhD: UCL, 1999.

[31] D. M. Binkley, B. J. Blalock, and J. M. Rochelle, "Optimizing drain current, inversion level, and channel length in analog CMOS design," *Analog Integrated Circuits and Signal Processing,* vol. 47, pp. 137-163, May 2006.

[32] A. I. A. Cunha, M. C. Schneider, and C. Galupmontoro, "An Explicit Physical Model for the Long-Channel Mos-Transistor Including Small-Signal Parameters," *Solid-State Electronics,* vol. 38, pp. 1945-1952, Nov 1995.

[33] D. M. Binkley, M. Bucher, and D. Foty, "Design-oriented characterization of CMOS over the continuum of inversion level and channel length," in *Electronics, Circuits and Systems, 2000. ICECS 2000. The 7th IEEE International Conference on*, 2000, pp. 161-164 vol.1.

[34] D. Flandre, A. Viviani, J. P. Eggermont, B. Gentinne, and P. G. A. Jespers, "Improved synthesis of gain-boosted regulated-cascode CMOS stages using symbolic analysis and gm/ID methodology," *Solid-State Circuits, IEEE Journal of,* vol. 32, pp. 1006-1012, 1997.

[35] F. Silveira, D. Flandre, and P. G. A. Jespers, "A gm-id based methodology for the design of CMOS analog circuits and its application to the synthesis of a silicon-on-insulator micropower OTA," *Solid-State Circuits, IEEE Journal of,* vol. 31, pp. 1314-1319, 1996.

[36] P. Aguirre, "Automatic Reusable Design for Analog Micropower Integrated Circuits." vol. Masters Montevideo, Uruguay: Instituto de Ingenieria Electrica, 2004.

[37] D. M. Binkley, C. E. Hopper, S. D. Tucker, B. C. Moss, J. M. Rochelle, and D. P. Foty, "A CAD methodology for optimizing transistor current and sizing in analog CMOS design," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on,* vol. 22, pp. 225-237, 2003.

[38] D. Foty, M. Bucher, and D. Binkley, "Re-interpreting the MOS transistor via the inversion coefficient and the continuum of g/sub ms//I/sub d," 2002, pp. 1179-1182 vol.3.

[39] C. Galup-Montoro, M. C. Schneider, and R. M. Coitinho, "Resizing rules for MOS analog-design reuse," *Design & Test of Computers, IEEE,* vol. 19, pp. 50-58, 2002.

[40] P. Giacomelli, M. C. Schneider, and C. Galup-Montoro, "MOSVIEW: a graphical tool for MOS analog design," 2003, pp. 45-46.

[41] D. S. Maher Kayal, "Procedural Analog Design," in *EKV Users' Meeting/Workshop*, Lausanne, 2004.

[42] "Analog Insydes http://www.analog-insydes.de/".

[43] "ISAAC http://homes.esat.kuleuven.be/~isaac/".

[44] B. Linares-Barranco and T. Serrano-Gotarredona, "On the design and characterization of femtoampere current-mode circuits," *Solid-State Circuits, IEEE Journal of,* vol. 38, pp. 1353-1363, 2003.

[45] D. R. S. Winton, "BSIM quote, Online: http://www.ece.msstate.edu/~winton/classes/ece4273/ ".

[46] W. Grabinski, "EKV Parameter Extraction Tutorial," in *IC-CAP Users' Conference*, Washington D.C., USA, 2001.

[47] C. C. Enz, F. Krummenacher, and E. A. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integrated Circuits and Signal Processing,* vol. 8, pp. 83-114, 1995.

[48] M. C. S. C. Galup-Montoro, Viriato C. Pahim, "Fundamentals of Next Generation Compact MOSFET Models," in *Proceedings of the 18th annual symposium on Integrated circuits and system design* Florianolpolis, France, 2005, pp. 32-37.

[49] J. M. Sallese, M. Bucher, F. Krummenacher, and P. Fazan, "Inversion charge linearization in MOSFET modeling and rigorous derivation of the EKV compact model," *Solid-State Electronics,* vol. 47, pp. 677-683, Apr 2003.

[50] S. C. Terry, J. M. Rochelle, D. M. Binkley, B. J. Blalock, D. P. Foty, and M. Bucher, "Comparison of a BSIM3V3 and EKV MOST model for a 0.5 um CMOS process and implications for analog circuit design," 2002, pp. 317-321 vol.1.

[51] G. Gildenblat, C. McAndrew, H. Wang, W. Wu, D. Foty, L. Lemaitre, and P. Bendix, "Advanced compact models: gateway to modern CMOS design," 2004, pp. 638-641.

[52] M. Bucher, C. Lallement, and C. C. Enz, "An efficient parameter extraction methodology for the EKV MOST model," in *Proceedings of International Conference on Microelectronic Test Structures. Trento, Italy. IEEE Electron Devices Soc. 25-28 March 1996.*, 1996.

[53] Z. Xunyu and C. Hutchens, "EKV model extraction for PD SOI MOSFET," in *Region 5 Conference: Annual Technical and Leadership Workshop, 2004*, 2004, pp. 81-83.

[54] M. K. Danica Stefanovic, "BSIM2EKV," in *EKV Users' Meeting/Workshop*, Lausanne, 2004.

[55] D. MacSweeney, K. G. McCarthy, A. Mathewson, and B. Mason, "A SPICE compatible subcircuit model for lateral bipolar transistors in a CMOS process," *Electron Devices, IEEE Transactions on,* vol. 45, pp. 1978-1984, 1998.

[56] E. A. Vittoz, "MOS transistors operated in the lateral bipolar mode and their application in CMOS technology," *Solid-State Circuits, IEEE Journal of,* vol. 18, pp. 273-279, 1983.

[57] V. S. Pershenkov, V. V. Belyakov, S. V. Cherepko, I. N. Shvetzov-Shilovsky, and V. V. Abramov, "Investigation of MOSFET operation in bipolar mode," 1997, pp. 273-276 vol.1.
[58] Y. Zhixin, M. J. Deen, and D. S. Malhi, "Gate-controlled lateral PNP BJT: characteristics, modeling and circuit applications," *Electron Devices, IEEE Transactions on,* vol. 44, pp. 118-128, 1997.
[59] S. G. Ravi Kumar N, Roy JN, Singh DN, "Design and realization of high performance CMOS compatible lateral bipolar transistors (CLBTs)," in *Proceedings of Spie - the International Society for Optical Engineering*, 2000, pp. 430-3.
[60] D. C. M. Corsi F, Marzocca C., "DC characterization of lateral bipolar devices in standard CMOS technology: a new model for base current partitioning," *Solid-State Electronics,* vol. vol.43, no.5, pp. 883-9, May 1999.
[61] R. R. Harrison and C. Charles, "A low-power low-noise CMOS amplifier for neural recording applications," *Ieee Journal of Solid-State Circuits,* vol. 38, pp. 958-965, Jun 2003.
[62] T. Matsuda, R. Minami, A. Kanamori, H. Iwata, T. Ohzone, S. Yamamoto, T. Ihara, and S. Nakajima, "A temperature and supply voltage independent CMOS voltage reference circuit," *Ieice Transactions on Electronics,* vol. E88C, pp. 1087-1093, May 2005.
[63] C. Liu, *Foundations of MEMS*. Upper Saddle River, NJ: Pearson/Prentice Hall, 2006.
[64] R. J. Baker, H. W. Li, D. E. Boyce, and Institute of Electrical and Electronics Engineers., *CMOS circuit design, layout, and simulation.* New York: IEEE Press, 1998.
[65] R. Rao, "Class Notes EE221 Advanced Semiconductor Devices, page 5-44," 2003.
[66] M. Shur, *Physics of semiconductor devices*. Englewood Cliffs, N.J.: Prentice Hall, 1990.
[67] B. G. Streetman and S. Banerjee, *Solid state electronic devices*, 5th ed. Upper Saddle River, N.J.: Prentice Hall, 2000.
[68] R. Pierret, *Field Effect Devices, Volume IV (2nd Edition)*: Addison-Wesley, January 1, 1990.
[69] R. Pierret, *Field Effect Devices, Volume IV (2nd Edition), page 135*: Addison-Wesley, January 1, 1990.
[70] D. Foty, "Perspectives on scaling theory and CMOS technology - understanding the past, present, and future," 2004, pp. 631-637.
[71] A. Hastings, *Art of Analog Layout, page 128*: Prentice Hall, 2000.
[72] A. A. Jaecklin, *Power semiconductor devices and circuits*. New York: Plenum Press, 1992.
[73] B. J. Baliga, *Power semiconductor devices*. Boston: PWS Pub. Co., 1996.

[74] V. e. Benda, J. Gowar, and D. A. Grant, *Power semiconductor devices : theory and applications*. Chichester ; New York: Wiley, 1999.
[75] A. S. S. a. K. C. Smith, *Microelectronic Circuits, Fourth Ed. (pp792-3)*. New York: Oxford University Press, 1998.
[76] R. Rao, "Class Notes EE221 Advanced Semiconductor Devices, page 9-38," 2003.
[77] E. S. Yang, *Microelectronic Devices, pp 292-3*. Boston Massachusetts: Kluwer, 1988.
[78] S. M. Sze, *Physics of semiconductor devices*, 2nd ed. New York: Wiley, 1981.
[79] A. Kranti, T. M. Chung, D. Flandre, and J. P. Raskin, "Laterally asymmetric channel engineering in fully depleted double gate SOI MOSFETs for high performance analog applications," *Solid-State Electronics,* vol. 48, pp. 947-959, Jun 2004.
[80] M. Hung, "Double Diffused (DMOS) FETs for Analog Applications," in *Electrical Engineering*. vol. PhD Boston, MA: MIT, 1991.
[81] J. P. Colinge, "Fully-depleted SOI CMOS for analog applications," *Electron Devices, IEEE Transactions on,* vol. 45, pp. 1010-1016, 1998.
[82] M. Stockinger, "Optimization of Ultra-Low-Power CMOS Transistors," in *Electrical Engineering*. vol. Ph.D.: Institut für Mikroelektronik, Eherfurt, Germany, 2000.
[83] E. Vittoz and J. Fellrath, "CMOS analog integrated circuits based on weak inversion operations," *Solid-State Circuits, IEEE Journal of,* vol. 12, pp. 224-231, 1977.
[84] C. Galup-Montoro, M. C. Schneider, and I. J. B. Loss, "Low output conductance composite MOSFET's for high frequency analog design," in *Circuits and Systems, 1994. ISCAS '94., 1994 IEEE International Symposium on*, 1994, pp. 783-786 vol.5.
[85] E. A. Vittoz, "Mos-Transistors Operated in the Lateral Bipolar Mode and Their Application in Cmos Technology," *Ieee Journal of Solid-State Circuits,* vol. 18, pp. 273-279, 1983.
[86] R. Gomez, R. Bashir, and G. W. Neudeck, "On the design and fabrication of novel lateral bipolar transistor in a deep-submicron technology," *Microelectronics Journal,* vol. 31, pp. 199-205, Mar 2000.
[87] M. Y. Hong, "Simulation and Fabrication of Submicron Channel-Length Dmos Transistors for Analog Applications," *IEEE Transactions on Electron Devices,* vol. 40, pp. 2222-2230, Dec 1993.